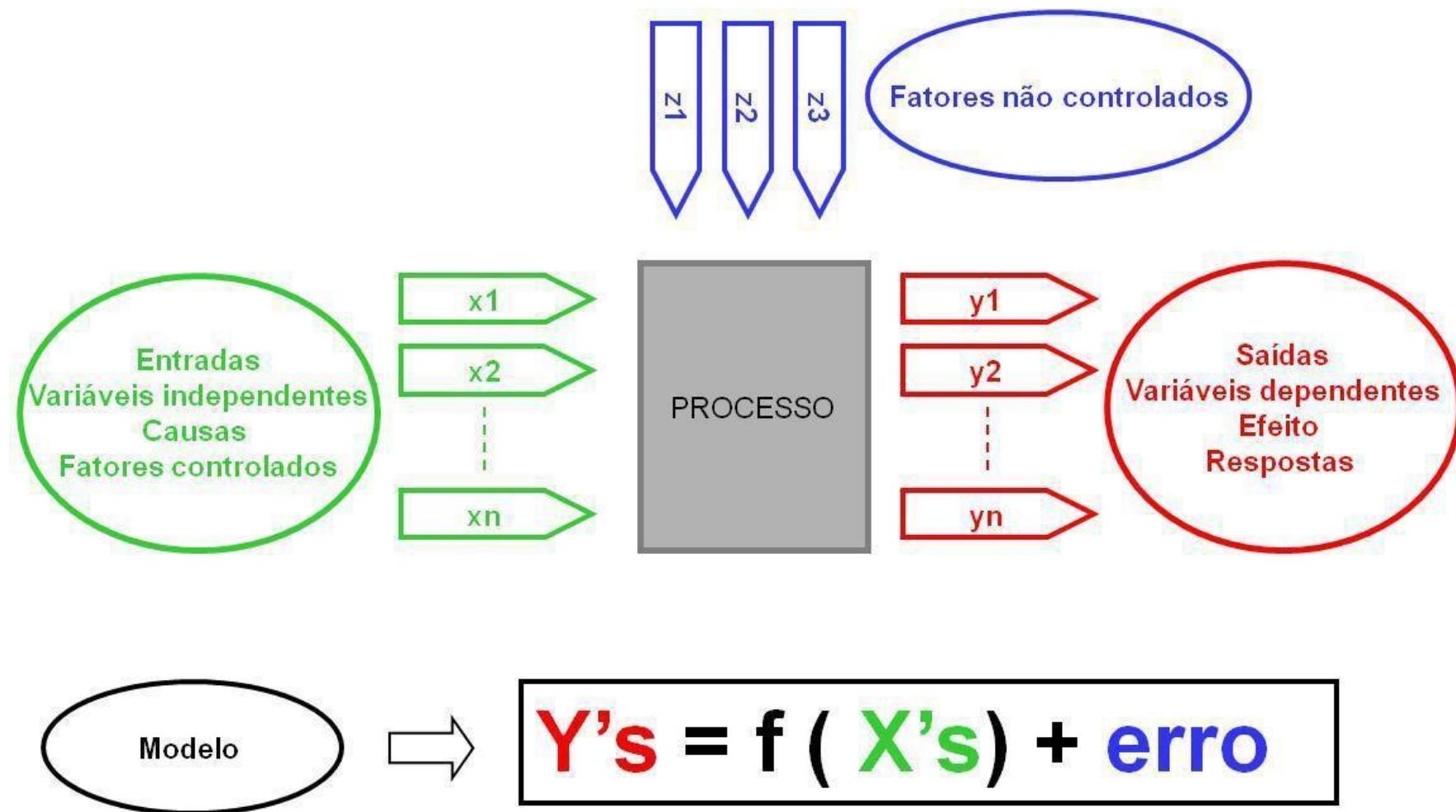


Variáveis do experimento



ERRO EXPERIMENTAL

- É percebido por quem realiza experimentos que duas unidades experimentais submetidas ao mesmo tratamento usualmente apresentam respostas diferentes. Essa variabilidade nas respostas é devido ao erro experimental.
- Através da análise estatística é possível medir a precisão das quantidades estimadas, como encontrados diferem significativamente de um valor pré estabelecido, ou ainda estabelecer intervalos de confiança para as quantidades estimadas. No contexto da estatística o termo erro se refere à variabilidade que ocorre nas respostas devido à atuação de fatores não controlados, embora possa incorporar erros grosseiros de procedimento.
- O erro experimental é devido a duas causas: a primeira delas é devido à variabilidade que existe entre as unidades experimentais e a segunda é devido à impossibilidade de se manter uniformes as condições de realização de um experimento.

VALOR DE “p”

Ele representa a probabilidade de ocorrência casual, provocada pela variabilidade normal do processo, de um dado resultado. Dessa forma um valor $p = 0,76$ indica que o resultado em questão tem 76% de chance de acontecer por acaso, assim como um valor $p = 0,003$ indica que o resultado tem somente 0,3 % de chance de acontecer por acaso.

• Nos estudos envolvendo processos industriais é comum adotar-se um nível de confiança de 95% (5% de significância). Esse é o nível de confiança que adotaremos em todos os testes durante o curso. A interpretação desse conceito deve ser feita da seguinte forma:

- qualquer alteração no resultado que apresente 5% ou mais de probabilidade (valor $p \geq 0,05$) de acontecer por acaso (pelas variações naturais do processo analisado) não deve ser atribuído ao fator analisado. Pode-se genericamente considerar que para valor $p \geq 0,05$ aquilo que se está analisando não tem diferença ou não tem influência.
- qualquer alteração no resultado que apresente menos de 5% probabilidade (valor $p < 0,05$) de acontecer por acaso, tem um valor tão baixo que é provável que ela tenha sido causada pela alteração no fator estudado e não pelo acaso. Quanto menor o valor p , maior a certeza dessa afirmação. Pode-se então genericamente considerar que para valor $p < 0,05$ aquilo que se está analisando é diferente ou tem influência.

1. Normalidade dos dados

Os métodos aqui tratados focam análises ditas paramétricas, para as quais se pressupõe uma distribuição normal dos valores analisados. As análises de dados não paramétricos (que não apresentam distribuição normal) não serão objeto desse curso.

A validade dos resultados obtidos através de testes paramétricos (baseados nos parâmetros media e variância da distribuição normal) é fortemente dependente da normalidade dos dados analisados. Salvo raras exceções, resultados de análises efetuadas por métodos paramétricos não são confiáveis se os dados não pertencem a amostras extraídas de populações normais. Por isso foram desenvolvidos testes para verificar a normalidade de distribuições de dados. Os mais difundidos são:

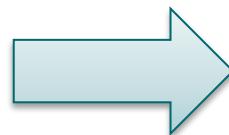
- Anderson-Darling,
- Shapiro-Wilk,
- Ryan-Joiner,
- Kolmogorov-Smirnov.

Teste de Hipótese

Objetivo: Verificar se os resultados experimentais obtidos a partir de determinada amostra contrariam ou não uma afirmação acerca de um parâmetro da população.

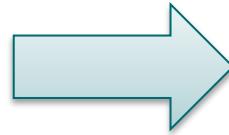
Imagine a situação em que um fabricante de determinado tipo de barra produzida atende a exigência de ter um comprimento médio de 70 cm.

Média populacional μ



Parâmetro de interesse da população

Produção de barras



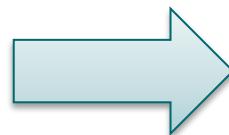
População

Teste de Hipótese

Objetivo: Verificar se os resultados experimentais obtidos a partir de determinada amostra contrariam ou não uma afirmação acerca de um parâmetro da população.

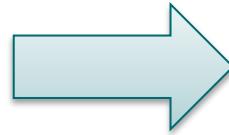
Imagine a situação em que um fabricante de determinado tipo de barra produzida atende a exigência de ter um comprimento médio de 70 cm.

Média populacional μ



Parâmetro de interesse da população

Produção de barras



População

Hipótese estatística ou hipótese nula: Afirmação ou conjectura sobre um parâmetro de interesse. Denotada por H_0 .

A hipótese nula é utilizada para verificar se a hipótese estabelecida inicialmente pode ser rejeitada ou não.

Hipótese alternativa: Hipótese utilizada como alternativa a hipótese nula. Denotada por H_1 .

Na situação anterior, temos que:

$$H_0: \mu = 70\text{cm}$$
$$H_1: \mu \neq 70\text{ cm}$$

Exemplo: Um gerente de produção está estudando a possibilidade de comprar uma nova máquina de estampar partes metálicas. Seja μ_0 o número médio de partes estampadas por hora pela máquina velha e μ a média da máquina nova. O gerente só comprará a máquina nova se ela for mais produtiva que a máquina velha.

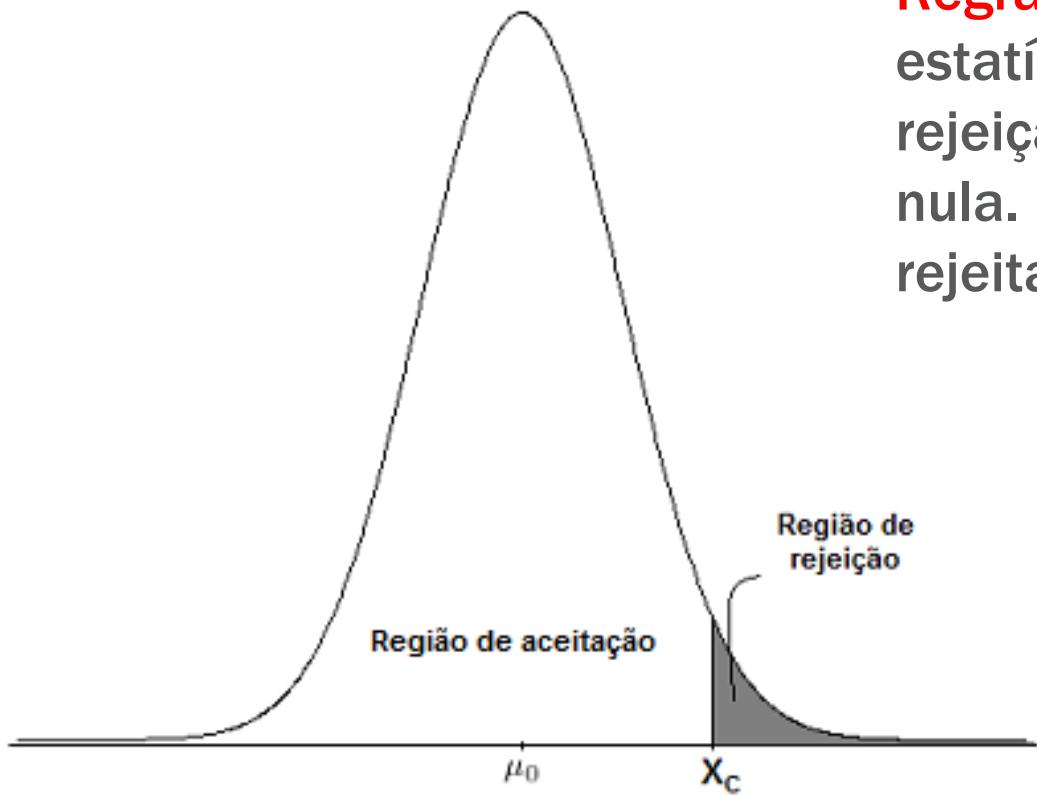
As hipóteses são definidas por:

$$\begin{aligned} H_0 &: \mu = \mu_0 \\ H_1 &: \mu > \mu_0 \end{aligned}$$

- Se a hipótese nula for rejeitada, então o gerente deve comprar a máquina nova
- Se a hipótese nula não for rejeitada, mantém-se a máquina antiga.

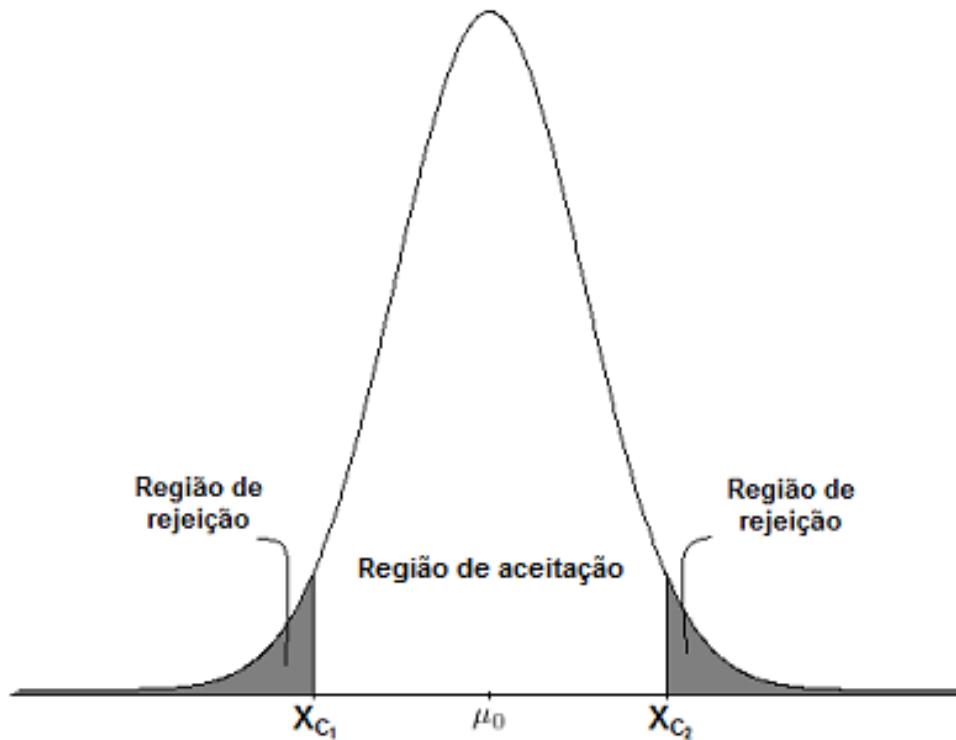
Região de Rejeição: A região de rejeição ou região crítica (R_C) é o conjunto de valores assumidos pela estatística de teste para os quais a hipótese nula é rejeitada. Seu complementar é a região de aceitação

Regra de Decisão: Se o valor da estatística pertence à região de rejeição, rejeitamos a hipótese nula. Caso contrário, não rejeitamos.



A região crítica é determinada segundo o teste a ser realizado. Além disso, a regra de decisão depende da hipótese alternativa H_1 . Por exemplo, para um teste bilateral, do tipo

$$H_0: \mu = \mu_0$$
$$H_1: \mu \neq \mu_0$$



Erros cometidos em testes de hipóteses

	Não rejeitar H_0	Rejeitar H_0
H_0 verdadeira	Decisão correta	Erro do tipo I
H_0 falsa	Erro do tipo II	Decisão correta

$$\alpha = P(\text{Erro do tipo I}) = P(\text{Rejeitar } H_0 | H_0 \text{ é verdadeira})$$

$$\beta = P(\text{Erro do tipo II}) = P(\text{Não rejeitar } H_0 | H_0 \text{ é falsa})$$

- **erro tipo 1:** Considerar como diferente algo que na realidade é igual (a probabilidade desse erro acontecer é igual ao nível de significância do teste, ou seja, 5% quando o nível de confiança é 95%).
 - **erro tipo 2:** Considerar como igual algo que na realidade é diferente.
- Não há como diminuir ao mesmo tempo os dois tipos de erro. Se diminuímos o erro tipo 1 (aumentando o nível de confiança) aumentamos a probabilidade de cometer um erro tipo 2 e vice versa.

Considere um teste unilateral do tipo:

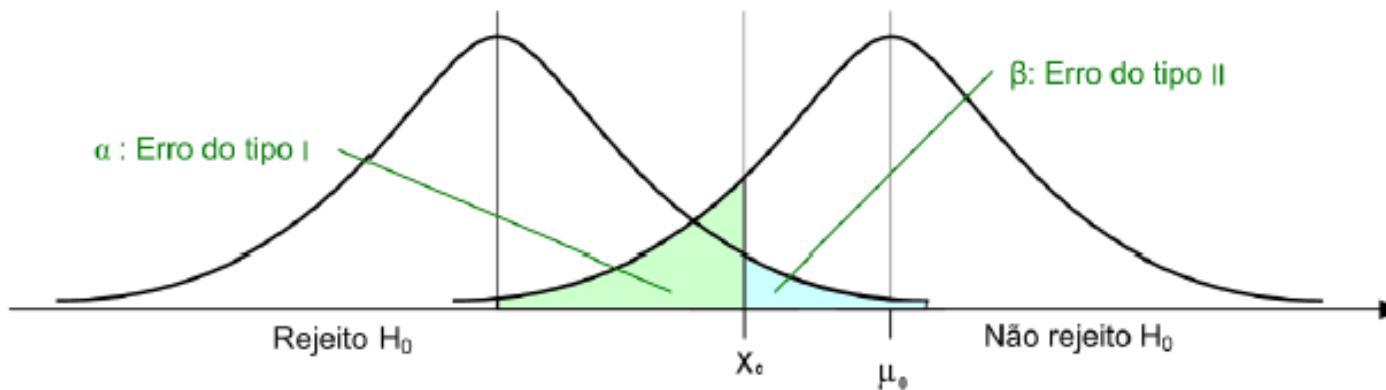
$$H_0: \mu = \mu_0$$

$$H_1: \mu < \mu_0$$

A região de rejeição, neste caso, é determinada por $\{\bar{X} < X_C\}$ e a interpretação dos erros pode ser vista da seguinte forma:

$$\alpha = P(\text{Erro do tipo I}) = P(\bar{X} < X_C | \mu = \mu_0)$$

$$\beta = P(\text{Erro do tipo II}) = P(\bar{X} > X_C | \mu < \mu_0)$$



Para testes de hipóteses como os considerados acima, podemos utilizar a seguinte expressão:

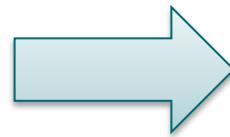
$$Z = \frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$$

que será a estatística do teste de hipóteses. Pelo teorema central do limite, sabemos que, se n é suficientemente grande, então esta estatística tem distribuição Normal padrão, ou seja,

$$Z \sim N(0,1)$$

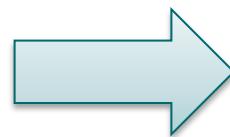
e, a partir dos valores de Z , podemos definir a região crítica.

$$\alpha = P(\text{Erro tipo I})$$



Nível de significância

$$1 - \alpha$$



Nível de confiança

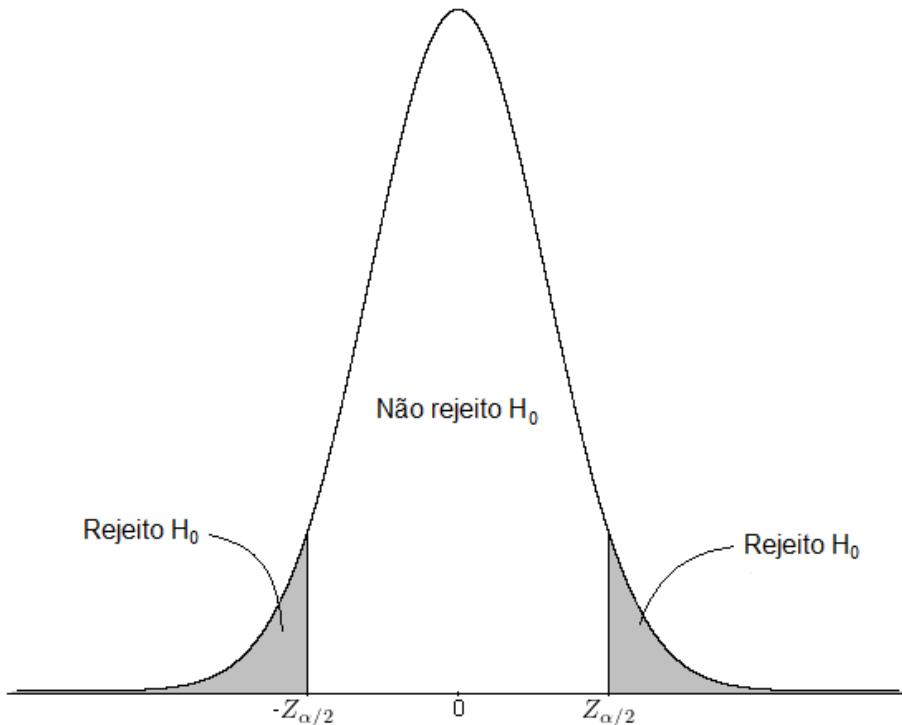
Considerando o nível de significância conhecido (α fixo) temos condições de determinar a região crítica (região de rejeição) do teste de hipóteses a partir da distribuição de Z.

Teste para Média – variância conhecida

1. Teste bilateral

$$H_0: \mu = \mu_0$$
$$H_1: \mu \neq \mu_0$$

Região crítica: teste bilateral



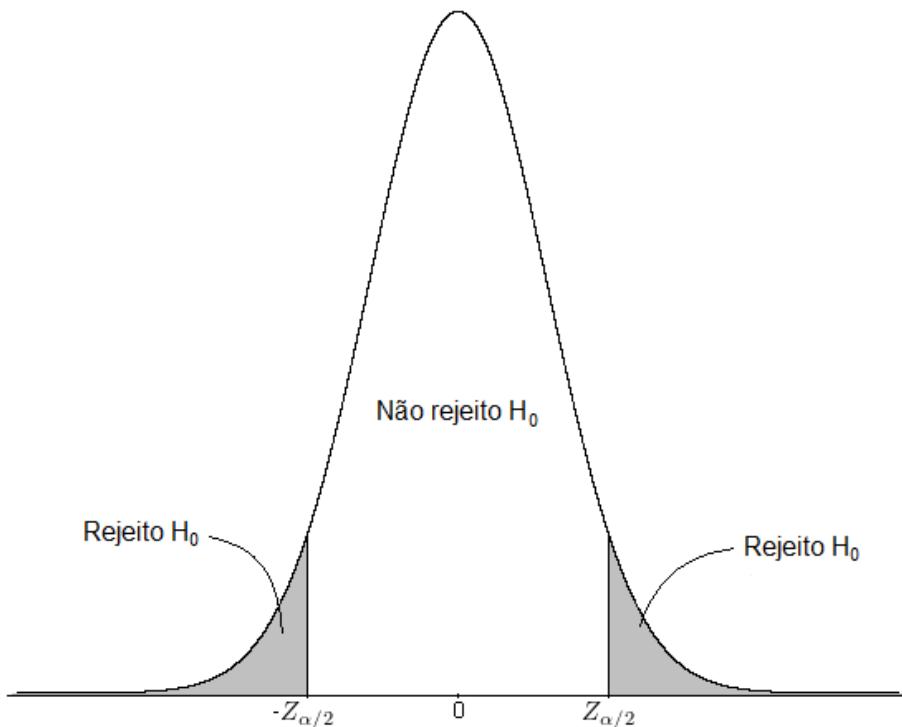
Pontos críticos:
 $-Z_{\alpha/2}$ e $Z_{\alpha/2}$

Teste para Média – variância conhecida

1. Teste bilateral

$$H_0: \mu = \mu_0$$
$$H_1: \mu \neq \mu_0$$

Região crítica: teste bilateral



Pontos críticos:
 $-Z_{\alpha/2}$ e $Z_{\alpha/2}$

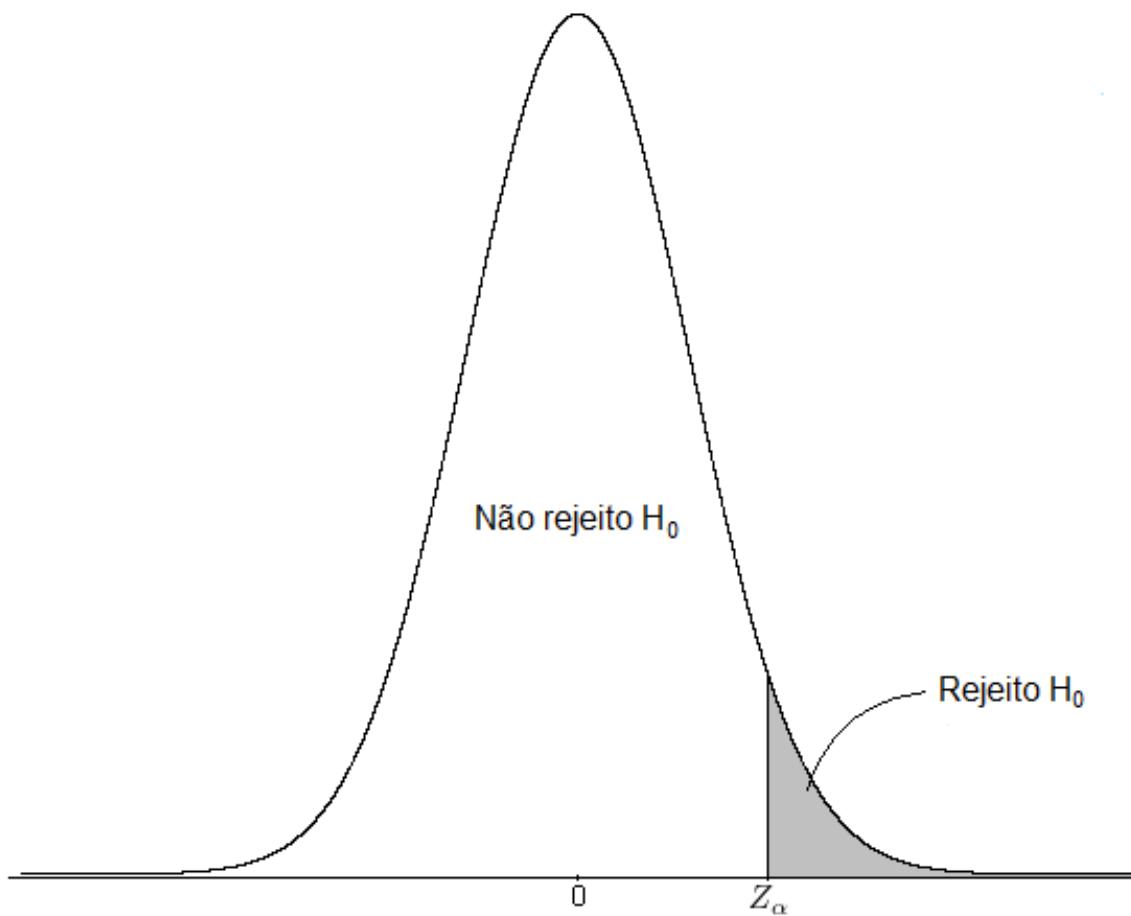
2. Teste unilateral à direita

$$H_0: \mu = \mu_0$$

$$H_1: \mu > \mu_0$$

Região crítica: teste unilateral à direita

Ponto crítico: Z_α

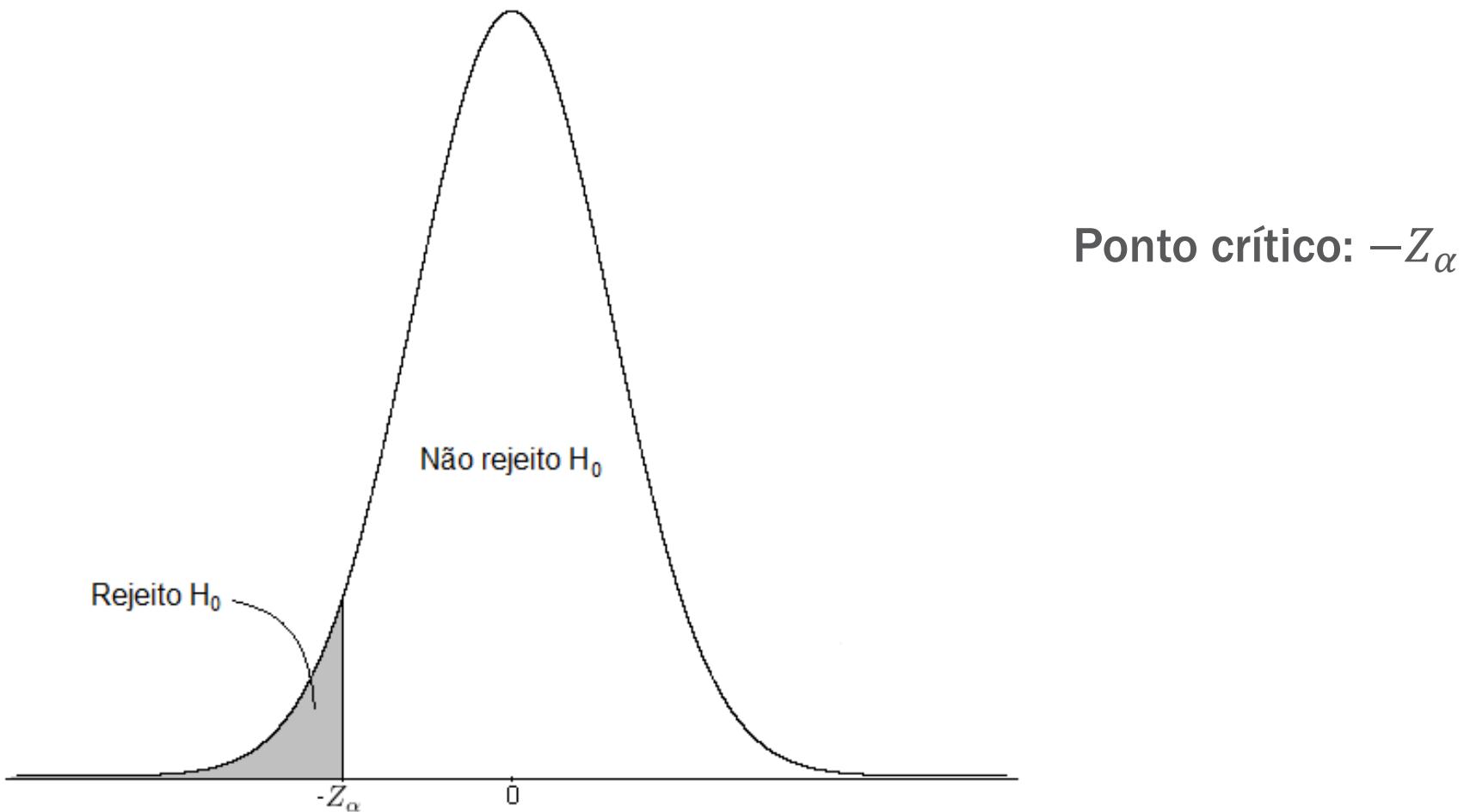


3. Teste unilateral à esquerda

$$H_0: \mu = \mu_0$$

$$H_1: \mu < \mu_0$$

Região crítica: teste unilateral à esquerda



Regra de Decisão

Hipótese alternativa	Rejeitamos H_0 se	Não rejeitamos H_0 se
$\mu \neq \mu_0$	$Z < -Z_{\alpha/2}$ ou se $Z > Z_{\alpha/2}$	$-Z_{\frac{\alpha}{2}} \leq Z \leq Z_{\alpha/2}$
$\mu > \mu_0$	$Z > Z_\alpha$	$Z \leq Z_\alpha$
$\mu < \mu_0$	$Z < -Z_\alpha$	$Z \geq -Z_\alpha$

Se o nível de significância é $\alpha = 5\%$, então $Z_\alpha = 1,64$ e $Z_{\alpha/2} = 1,96$.

Exemplo 5.1.1.1: Um supervisor da qualidade quer testar, com base em uma amostra aleatória de tamanho $n = 35$ e, para um nível de significância $\alpha = 5\%$ se a profundidade média de um furo numa determinada peça é $72,4 \text{ mm}$. O que podemos dizer se ele obteve $\bar{X} = 73,2 \text{ mm}$ e se sabe, de informações anteriores, que $\sigma = 2,1 \text{ mm}$?

1. Inicialmente, estabelecemos as hipóteses

$$H_0: \mu = 72,4 \text{ mm}$$

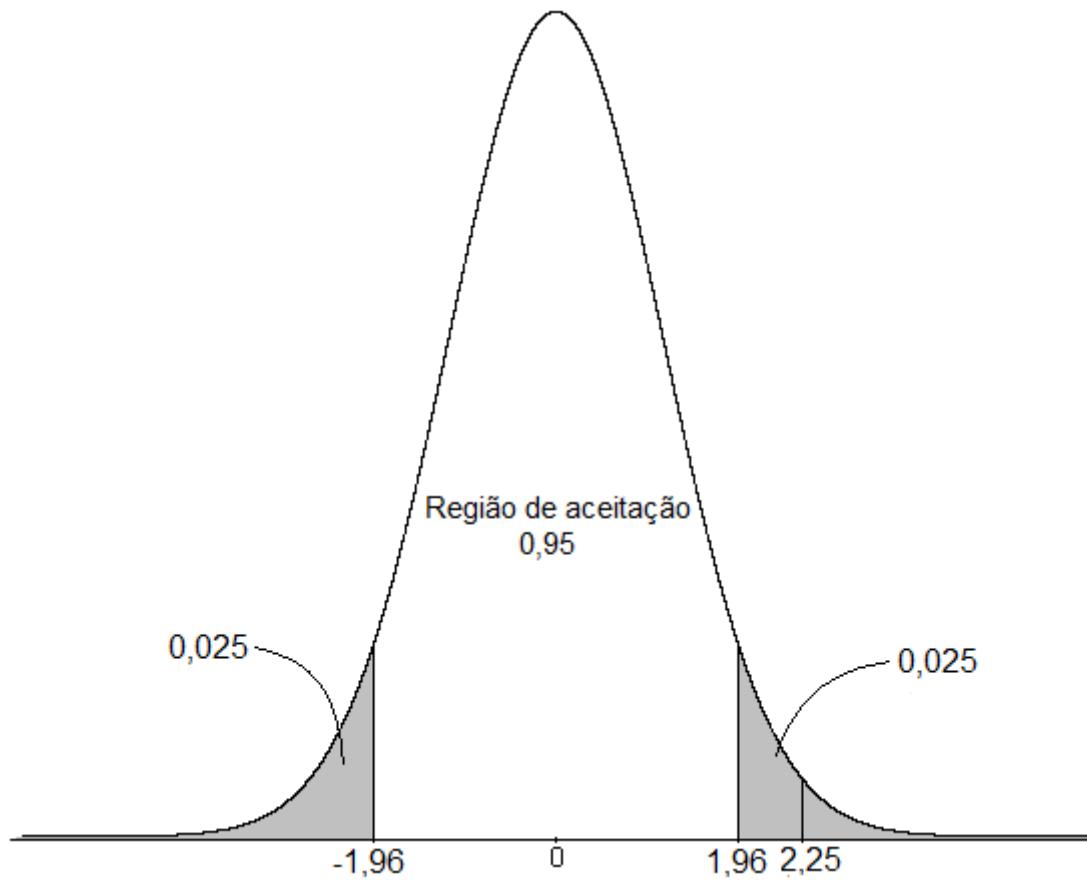
$$H_1: \mu \neq 72,4 \text{ mm}$$

2. Como $\alpha = 0,05$, temos que $Z_{\alpha/2} = Z_{0,025} = 1,96$.

3. Calculamos o valor da estatística

$$Z_{obs} = \frac{\bar{X} - 72,4}{\frac{2,1}{\sqrt{35}}} = \frac{73,2 - 72,4}{\frac{2,1}{\sqrt{35}}} = 2,25$$

4. Regra de decisão: Rejeitamos H_0 se $Z_{obs} > 1,96$ ou se $Z_{obs} < -1,96$.



Conclusão: Como $Z_{obs} = 2,25 > 1,96$ rejeitamos a hipótese nula de que a profundidade média do furo na peça é de 72,4 mm.

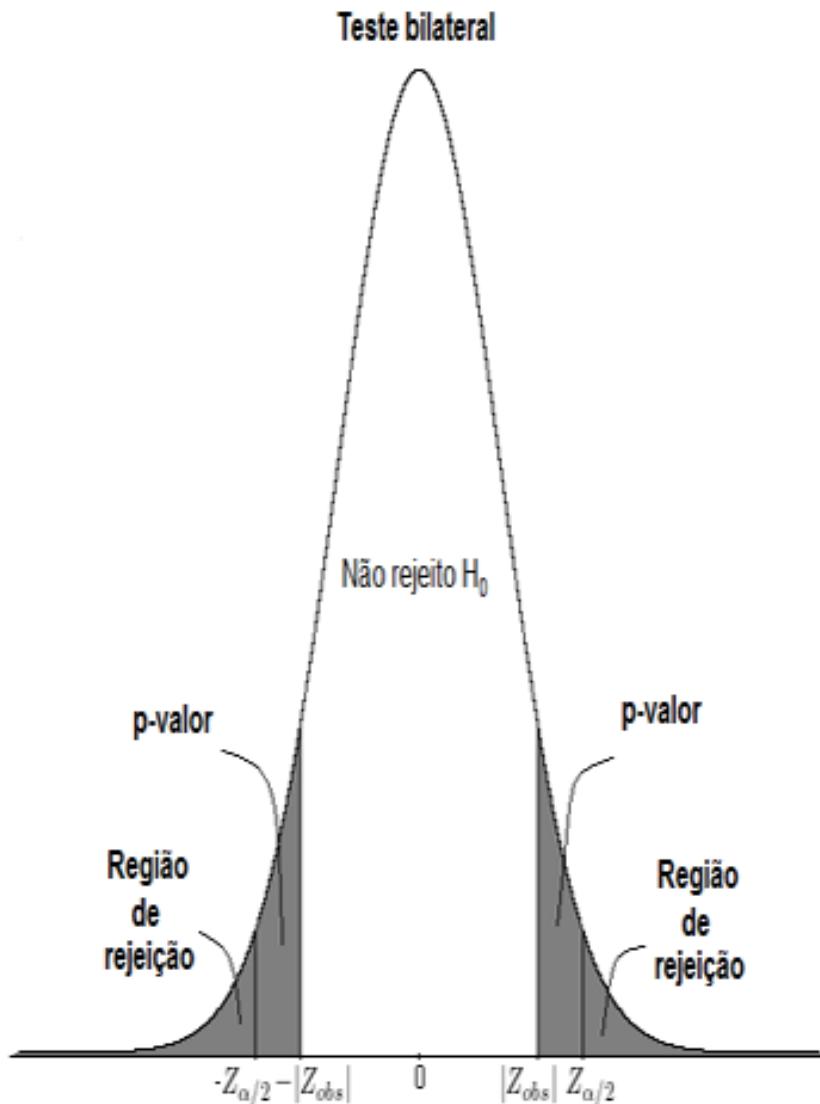
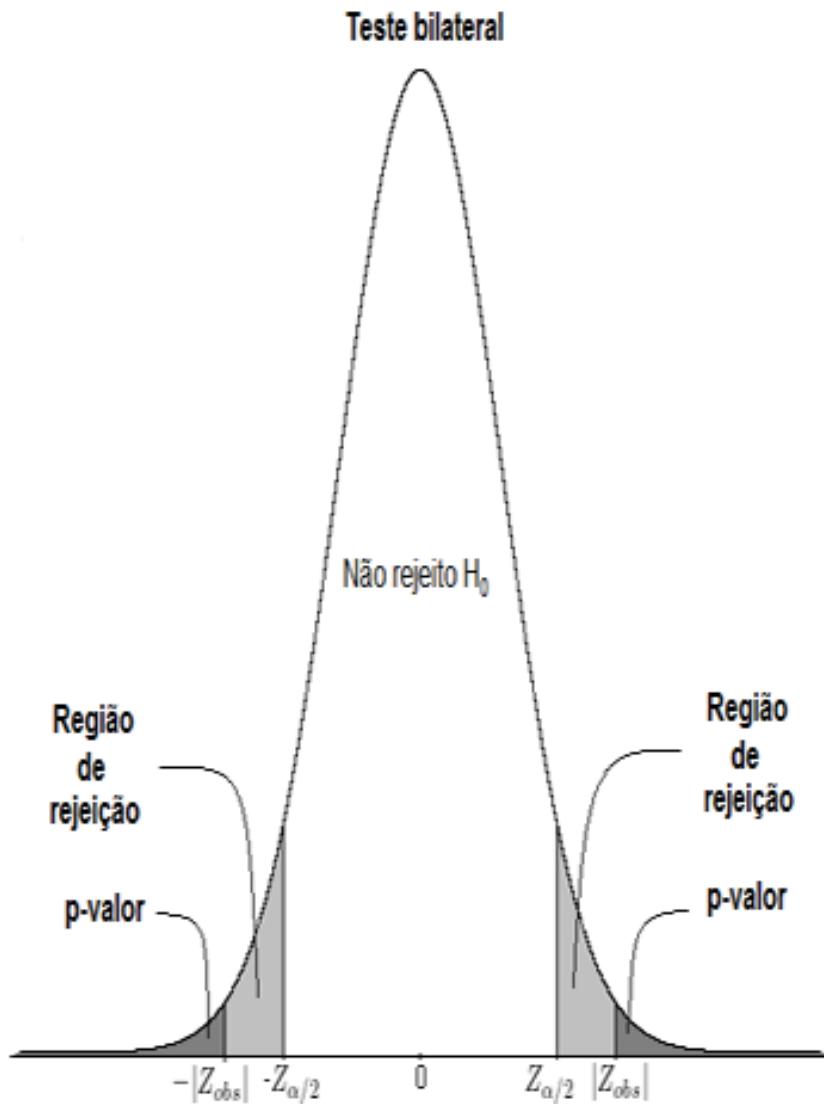
Cálculo e Interpretação do p-valor

Definição:

O p-valor é a probabilidade de se obter uma estatística de teste igual ou mais extrema que aquela observada em uma amostra, sob a hipótese nula.

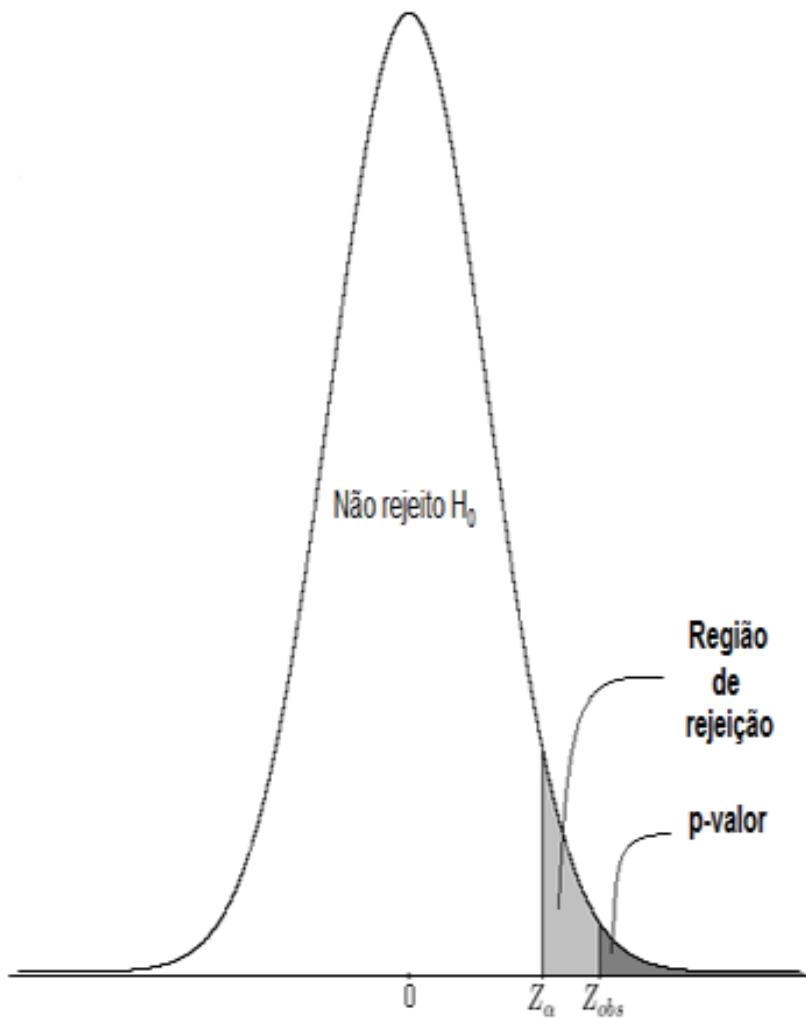
- É o menor nível de significância para o qual rejeitamos a hipótese nula.
- Um p-valor pequeno significa que a probabilidade de obter um valor da estatística de teste como o observado é muito improvável, consequentemente, rejeita-se a hipótese nula.

Teste bilateral

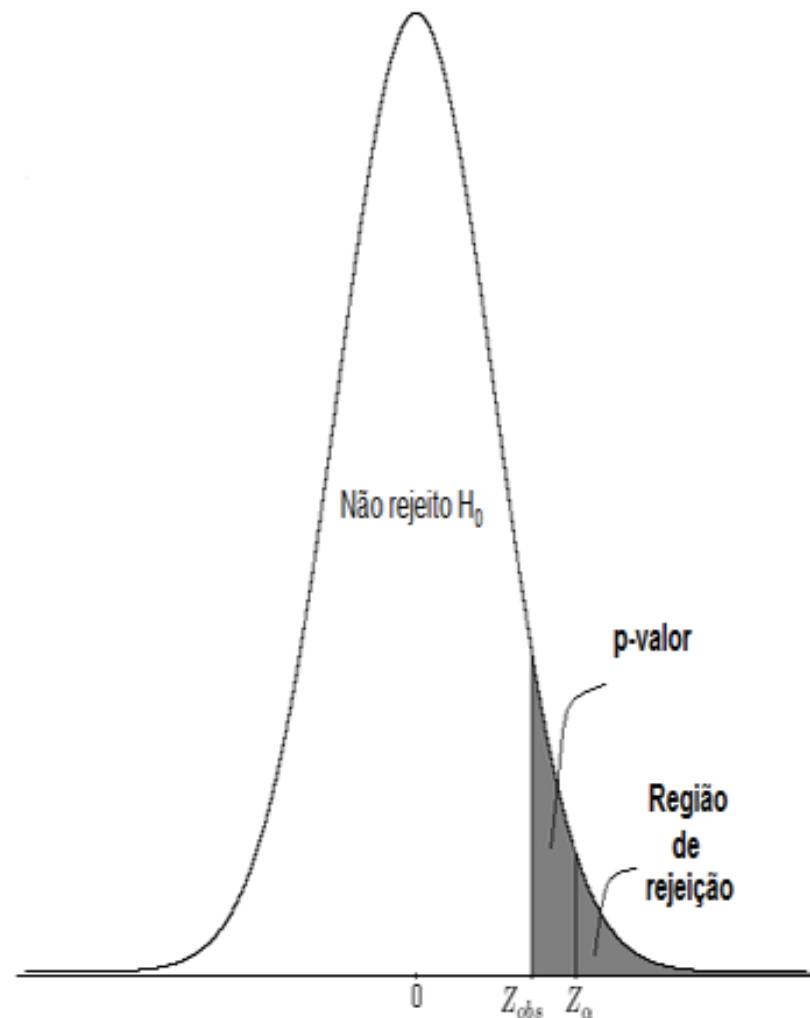


Teste unilateral

Teste unilateral à direita



Teste unilateral à direita



Observação: Se o p-valor é menor que o nível de significância proposto, então o valor da estatística Z_{obs} estará na região crítica e, portanto, rejeitamos a hipótese nula H_0 . Em contrapartida, se o p-valor é maior que o nível de significância, não rejeitamos a hipótese nula H_0 .

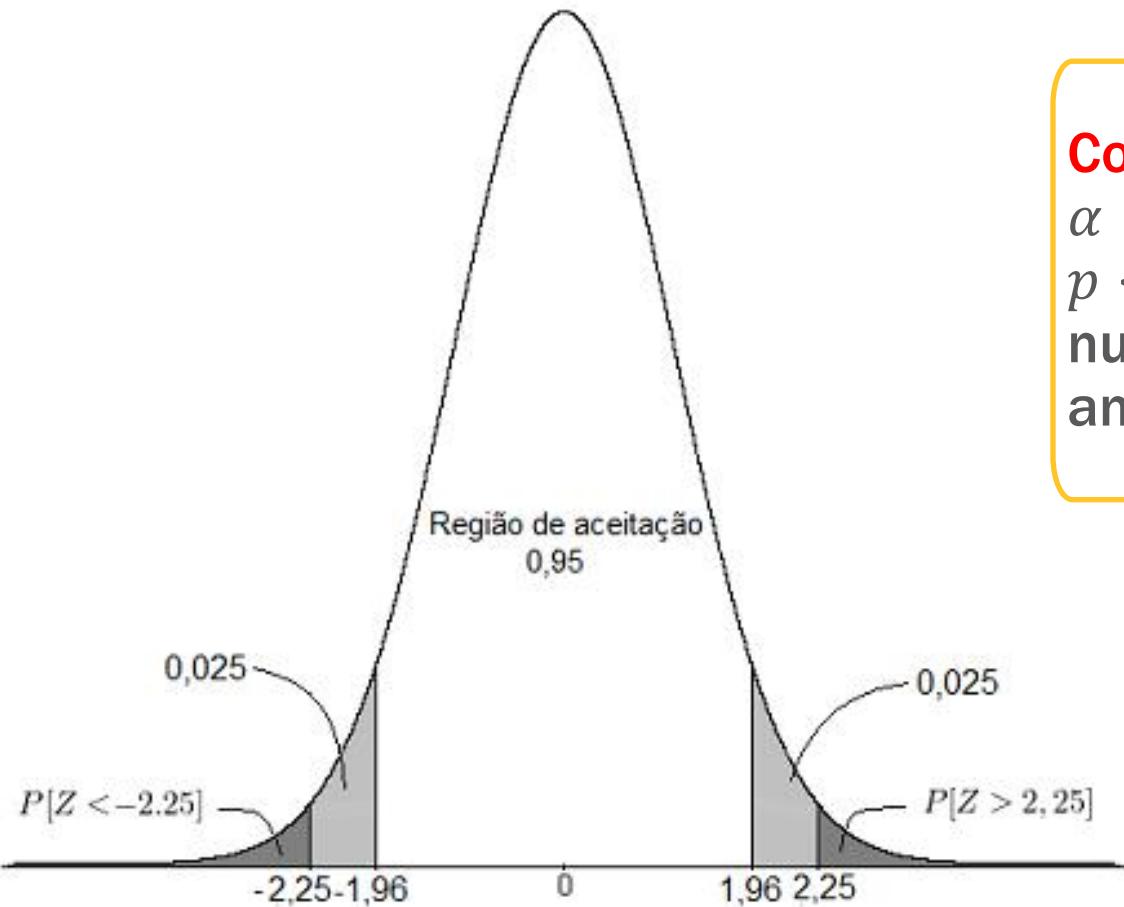
Exemplo 5.1.1.1: Um supervisor da qualidade quer testar, com base em uma amostra aleatória de tamanho $n = 35$ e, para um nível de significância $\alpha = 5\%$ se a profundidade média de um furo numa determinada peça é $72,4\text{ mm}$. O que podemos dizer se ele obteve $\bar{X} = 73,2\text{ mm}$ e se sabe, de informações anteriores, que $\sigma = 2,1\text{ mm}$?

Já vimos, neste exemplo, que o valor da estatística Z_{obs} é dado por:

$$Z_{obs} = 2,25$$

Desta forma, temos que $|Z_{obs}| = 2,25$ e, como o teste é bilateral, temos que o p-valor é dado por

$$p = 2 \times P(Z > |Z_{obs}|) = 2 \times P(Z > 2,25) = 0,0244$$



Conclusão: Considerando $\alpha = 0,05$, como o p-valor $p < \alpha$, rejeitamos a hipótese nula. Como já sabíamos anteriormente.

Hipótese estatística ou hipótese nula: Afirmação ou conjectura sobre um parâmetro de interesse. Denotada por H_0 .

A hipótese nula é utilizada para verificar se a hipótese estabelecida inicialmente pode ser rejeitada ou não.

Hipótese alternativa: Hipótese utilizada como alternativa a hipótese nula. Denotada por H_1 .

Na situação anterior, temos que:

$$H_0: \mu = 70\text{cm}$$
$$H_1: \mu \neq 70\text{ cm}$$

Exemplo: Um gerente de produção está estudando a possibilidade de comprar uma nova máquina de estampar partes metálicas. Seja μ_0 o número médio de partes estampadas por hora pela máquina velha e μ a média da máquina nova. O gerente só comprará a máquina nova se ela for mais produtiva que a máquina velha.

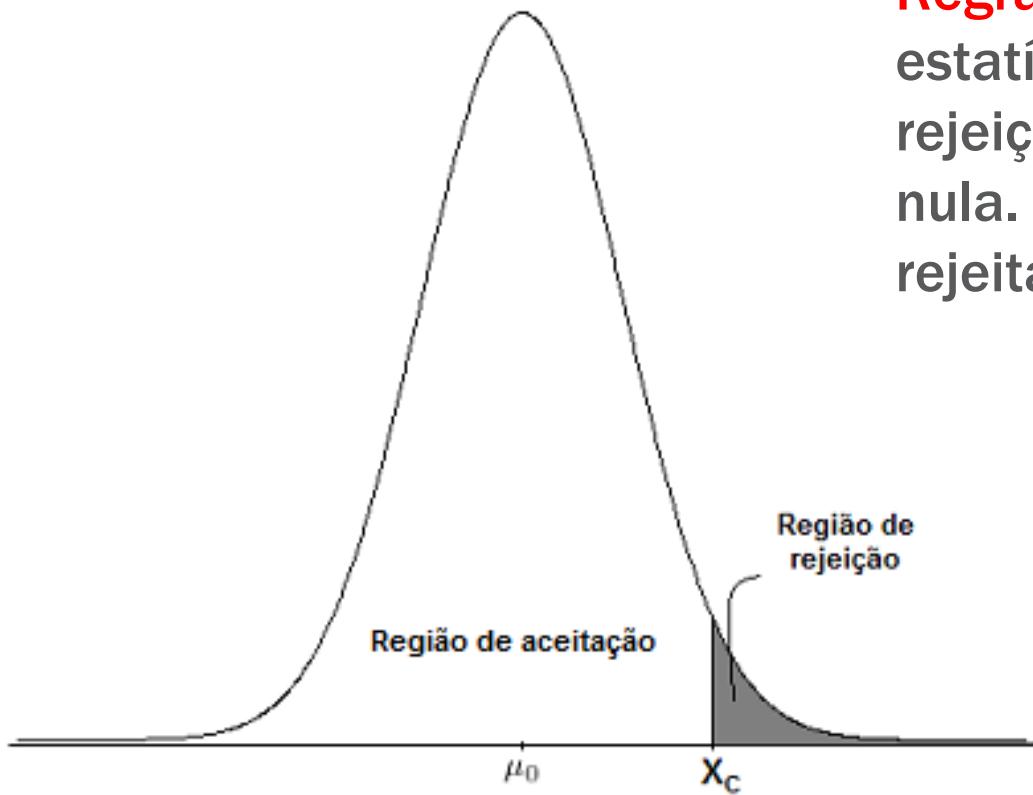
As hipóteses são definidas por:

$$\begin{aligned} H_0 &: \mu = \mu_0 \\ H_1 &: \mu > \mu_0 \end{aligned}$$

- Se a hipótese nula for rejeitada, então o gerente deve comprar a máquina nova
- Se a hipótese nula não for rejeitada, mantém-se a máquina antiga.

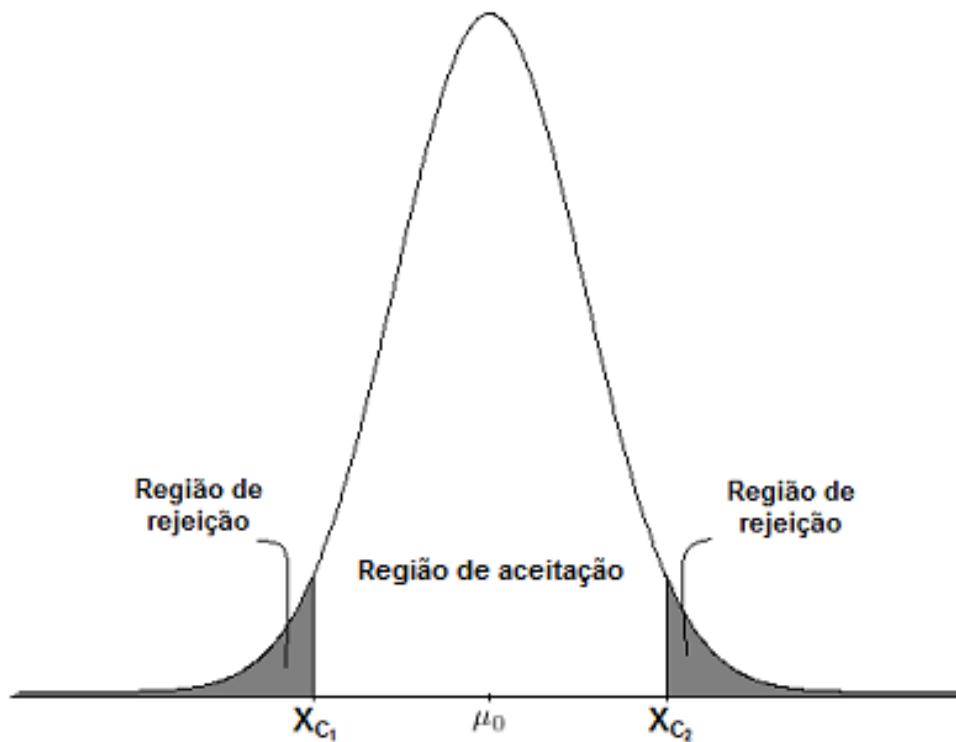
Região de Rejeição: A região de rejeição ou região crítica (R_C) é o conjunto de valores assumidos pela estatística de teste para os quais a hipótese nula é rejeitada. Seu complementar é a região de aceitação

Regra de Decisão: Se o valor da estatística pertence à região de rejeição, rejeitamos a hipótese nula. Caso contrário, não rejeitamos.



A região crítica é determinada segundo o teste a ser realizado. Além disso, a regra de decisão depende da hipótese alternativa H_1 . Por exemplo, para um teste bilateral, do tipo

$$H_0: \mu = \mu_0$$
$$H_1: \mu \neq \mu_0$$



Erros cometidos em testes de hipóteses

	Não rejeitar H_0	Rejeitar H_0
H_0 verdadeira	Decisão correta	Erro do tipo I
H_0 falsa	Erro do tipo II	Decisão correta

$$\alpha = P(\text{Erro do tipo I}) = P(\text{Rejeitar } H_0 | H_0 \text{ é verdadeira})$$

$$\beta = P(\text{Erro do tipo II}) = P(\text{Não rejeitar } H_0 | H_0 \text{ é falsa})$$

Considere um teste unilateral do tipo:

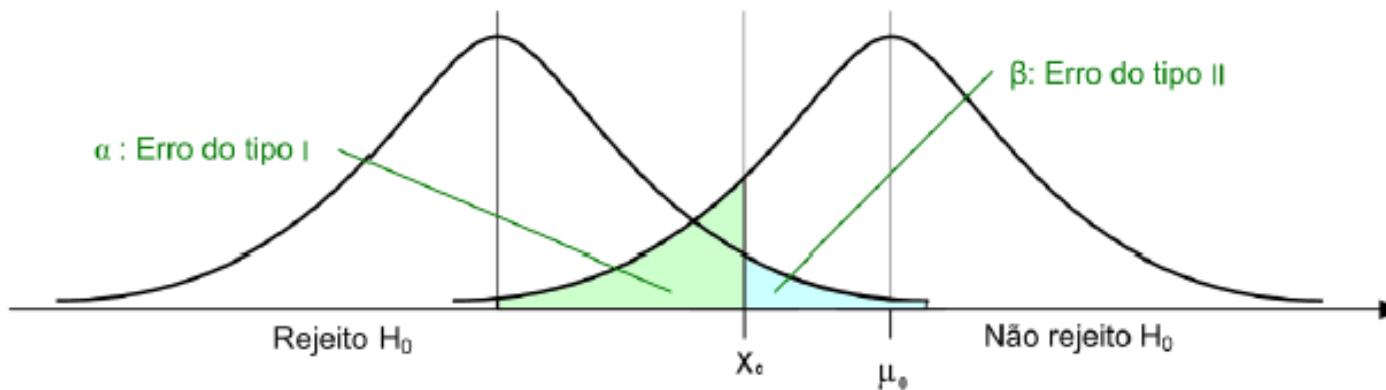
$$H_0: \mu = \mu_0$$

$$H_1: \mu < \mu_0$$

A região de rejeição, neste caso, é determinada por $\{\bar{X} < X_C\}$ e a interpretação dos erros pode ser vista da seguinte forma:

$$\alpha = P(\text{Erro do tipo I}) = P(\bar{X} < X_C | \mu = \mu_0)$$

$$\beta = P(\text{Erro do tipo II}) = P(\bar{X} > X_C | \mu < \mu_0)$$



Para testes de hipóteses como os considerados acima, podemos utilizar a seguinte expressão:

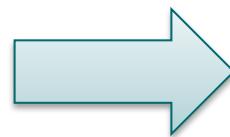
$$Z = \frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$$

que será a estatística do teste de hipóteses. Pelo teorema central do limite, sabemos que, se n é suficientemente grande, então esta estatística tem distribuição Normal padrão, ou seja,

$$Z \sim N(0,1)$$

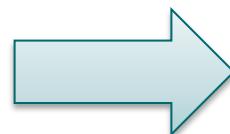
e, a partir dos valores de Z , podemos definir a região crítica.

$$\alpha = P(\text{Erro tipo I})$$



Nível de significância

$$1 - \alpha$$



Nível de confiança

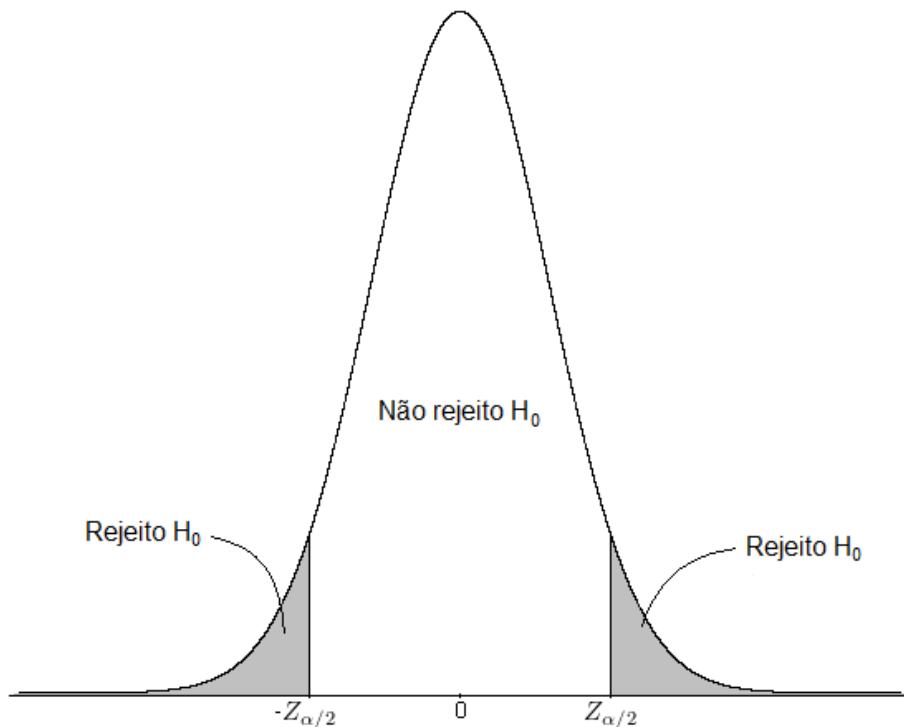
Considerando o nível de significância conhecido (α fixo) temos condições de determinar a região crítica (região de rejeição) do teste de hipóteses a partir da distribuição de Z.

Teste para Média – variância conhecida

1. Teste bilateral

$$H_0: \mu = \mu_0$$
$$H_1: \mu \neq \mu_0$$

Região crítica: teste bilateral

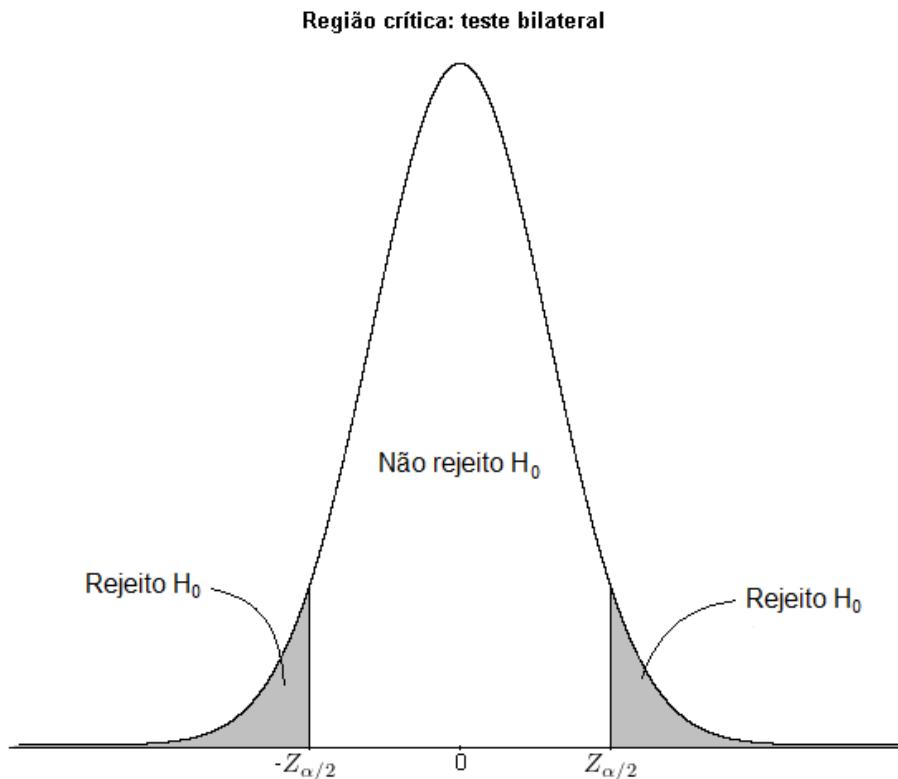


Pontos críticos:
 $-Z_{\alpha/2}$ e $Z_{\alpha/2}$

Teste para Média – variância conhecida

1. Teste bilateral

$$H_0: \mu = \mu_0$$
$$H_1: \mu \neq \mu_0$$



Pontos críticos:
 $-Z_{\alpha/2}$ e $Z_{\alpha/2}$

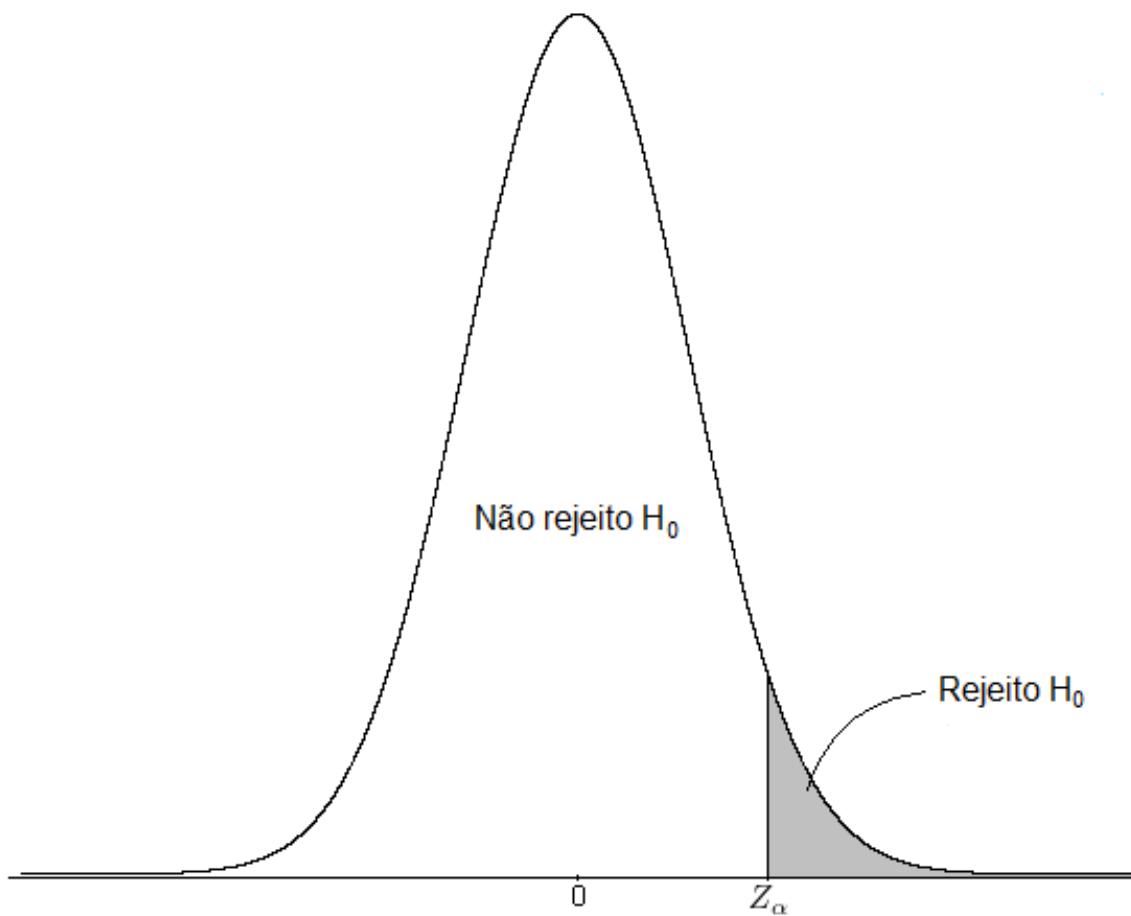
2. Teste unilateral à direita

$$H_0: \mu = \mu_0$$

$$H_1: \mu > \mu_0$$

Região crítica: teste unilateral à direita

Ponto crítico: Z_α

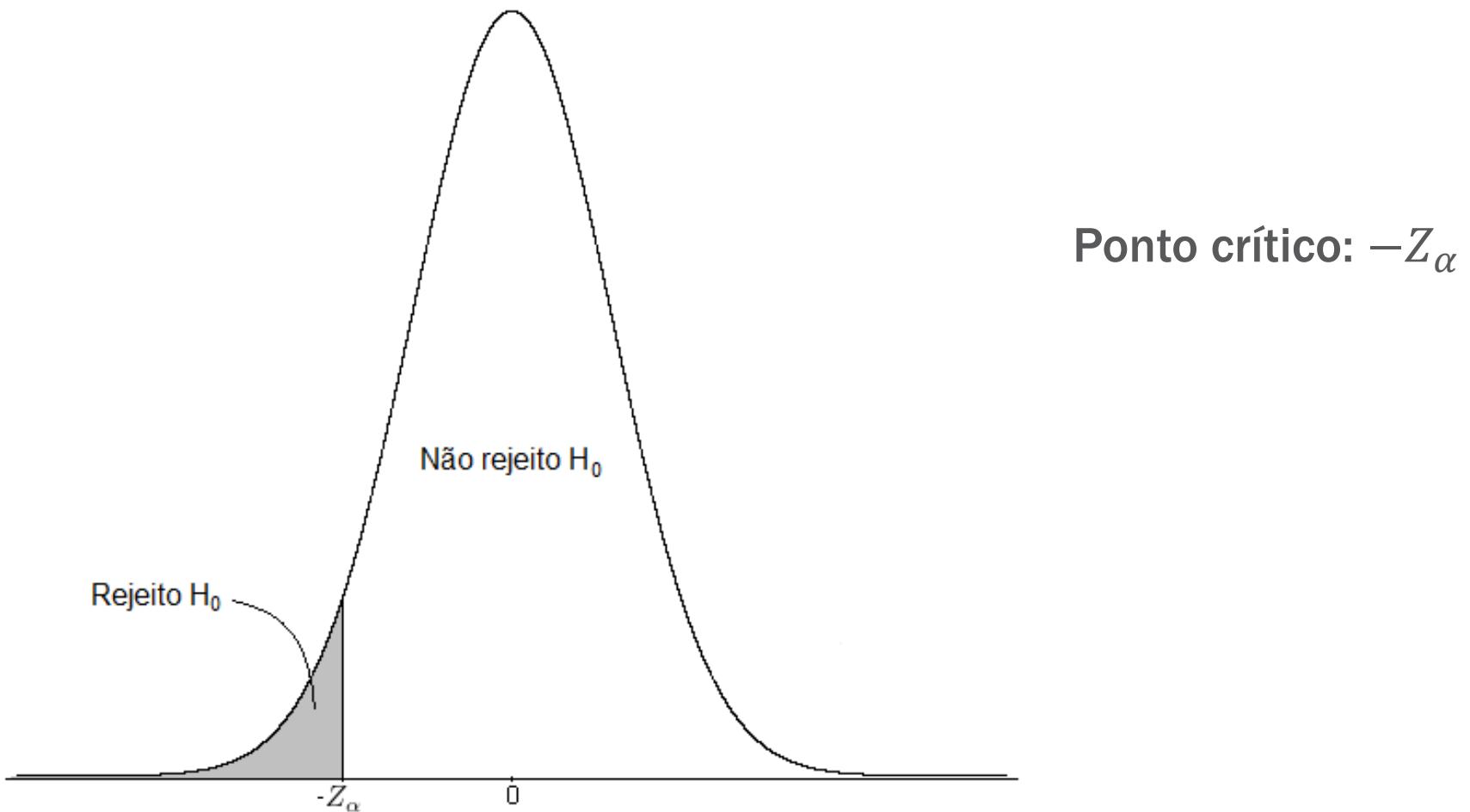


3. Teste unilateral à esquerda

$$H_0: \mu = \mu_0$$

$$H_1: \mu < \mu_0$$

Região crítica: teste unilateral à esquerda



Regra de Decisão

Hipótese alternativa	Rejeitamos H_0 se	Não rejeitamos H_0 se
$\mu \neq \mu_0$	$Z < -Z_{\alpha/2}$ ou se $Z > Z_{\alpha/2}$	$-Z_{\frac{\alpha}{2}} \leq Z \leq Z_{\alpha/2}$
$\mu > \mu_0$	$Z > Z_\alpha$	$Z \leq Z_\alpha$
$\mu < \mu_0$	$Z < -Z_\alpha$	$Z \geq -Z_\alpha$

Se o nível de significância é $\alpha = 5\%$, então $Z_\alpha = 1,64$ e $Z_{\alpha/2} = 1,96$.

Exemplo 5.1.1.1: Um supervisor da qualidade quer testar, com base em uma amostra aleatória de tamanho $n = 35$ e, para um nível de significância $\alpha = 5\%$ se a profundidade média de um furo numa determinada peça é $72,4 \text{ mm}$. O que podemos dizer se ele obteve $\bar{X} = 73,2 \text{ mm}$ e se sabe, de informações anteriores, que $\sigma = 2,1 \text{ mm}$?

1. Inicialmente, estabelecemos as hipóteses

$$H_0: \mu = 72,4 \text{ mm}$$

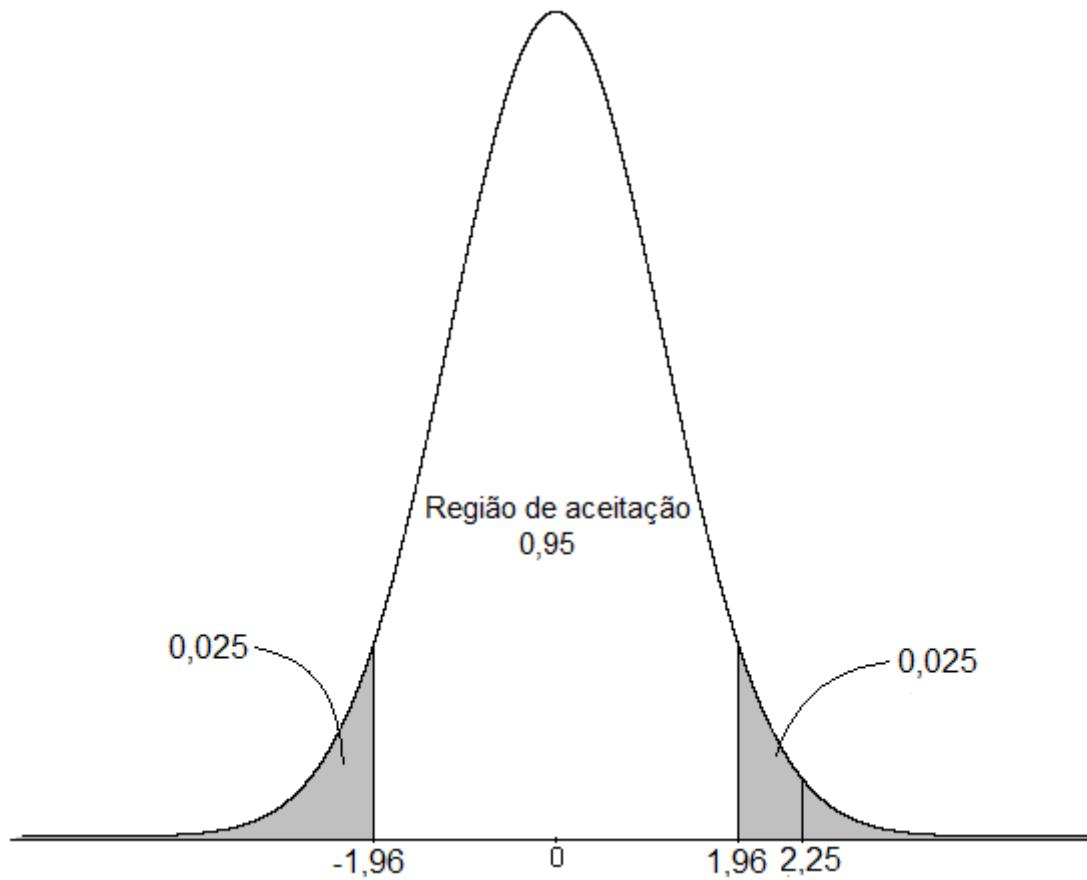
$$H_1: \mu \neq 72,4 \text{ mm}$$

2. Como $\alpha = 0,05$, temos que $Z_{\alpha/2} = Z_{0,025} = 1,96$.

3. Calculamos o valor da estatística

$$Z_{obs} = \frac{\bar{X} - 72,4}{\frac{2,1}{\sqrt{35}}} = \frac{73,2 - 72,4}{\frac{2,1}{\sqrt{35}}} = 2,25$$

4. Regra de decisão: Rejeitamos H_0 se $Z_{obs} > 1,96$ ou se $Z_{obs} < -1,96$.



Conclusão: Como $Z_{obs} = 2,25 > 1,96$ rejeitamos a hipótese nula de que a profundidade média do furo na peça é de 72,4 mm.

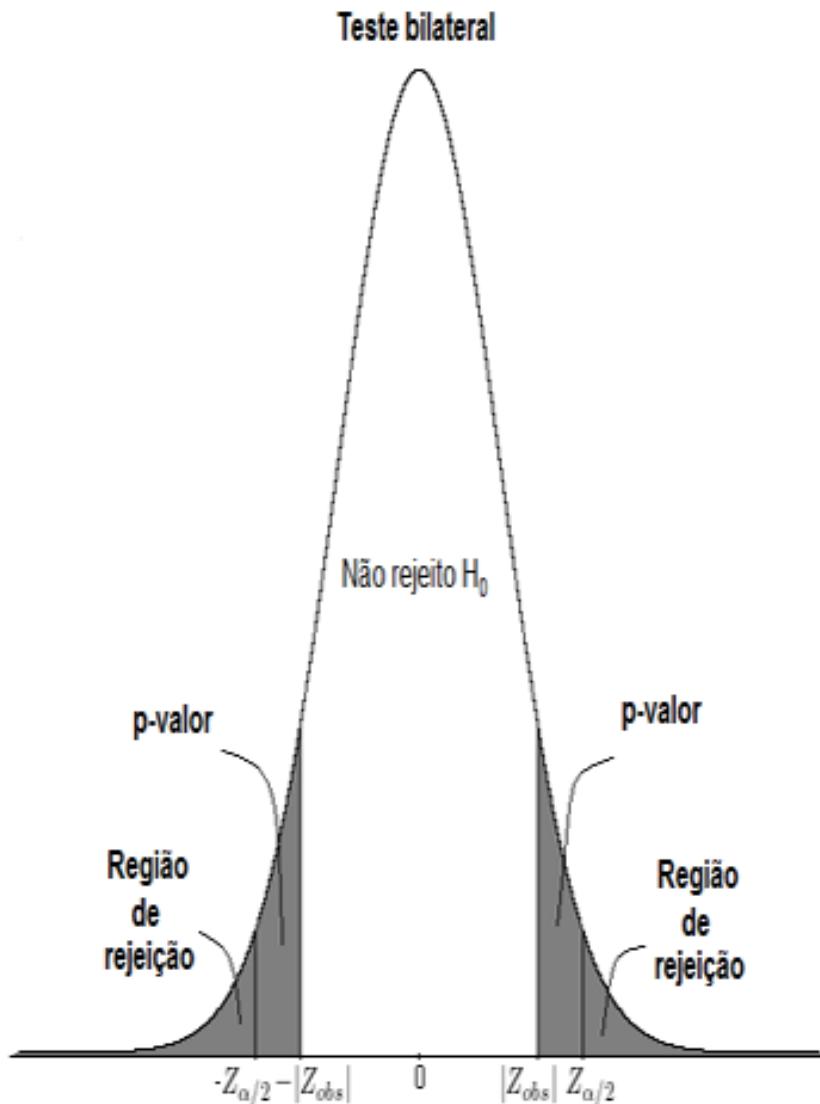
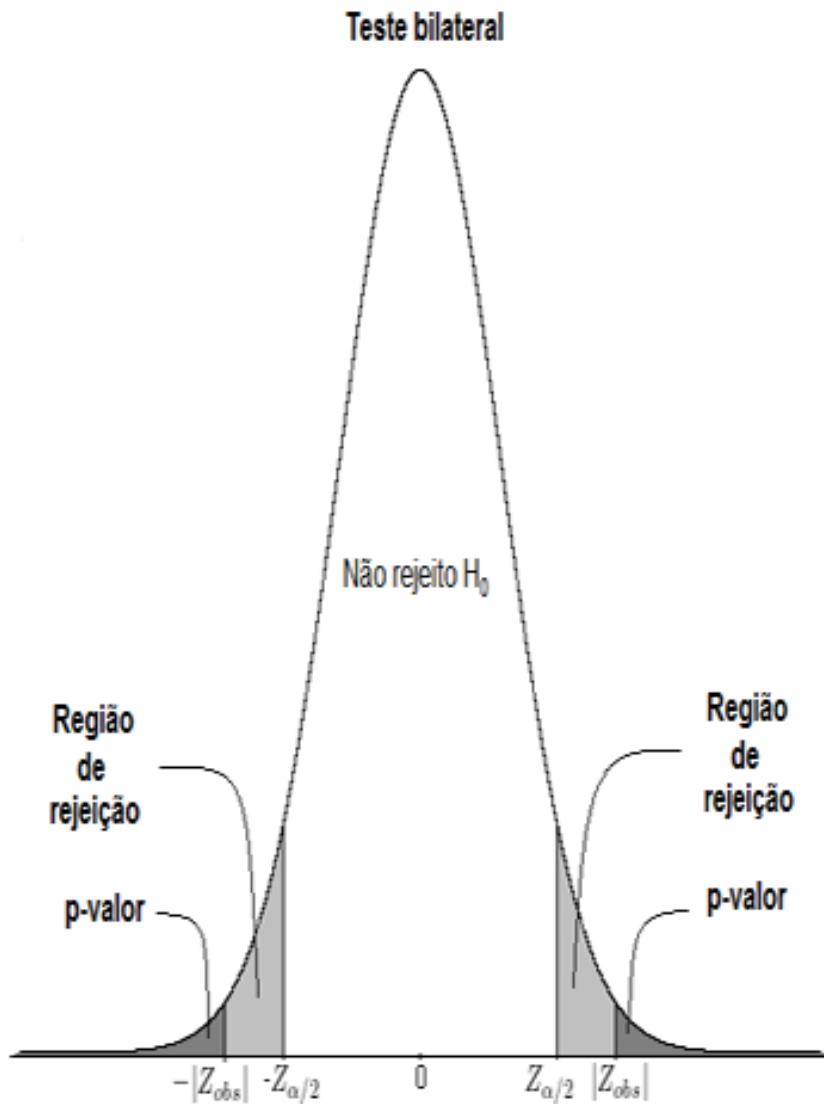
Cálculo e Interpretação do p-valor

Definição:

O p-valor é a probabilidade de se obter uma estatística de teste igual ou mais extrema que aquela observada em uma amostra, sob a hipótese nula.

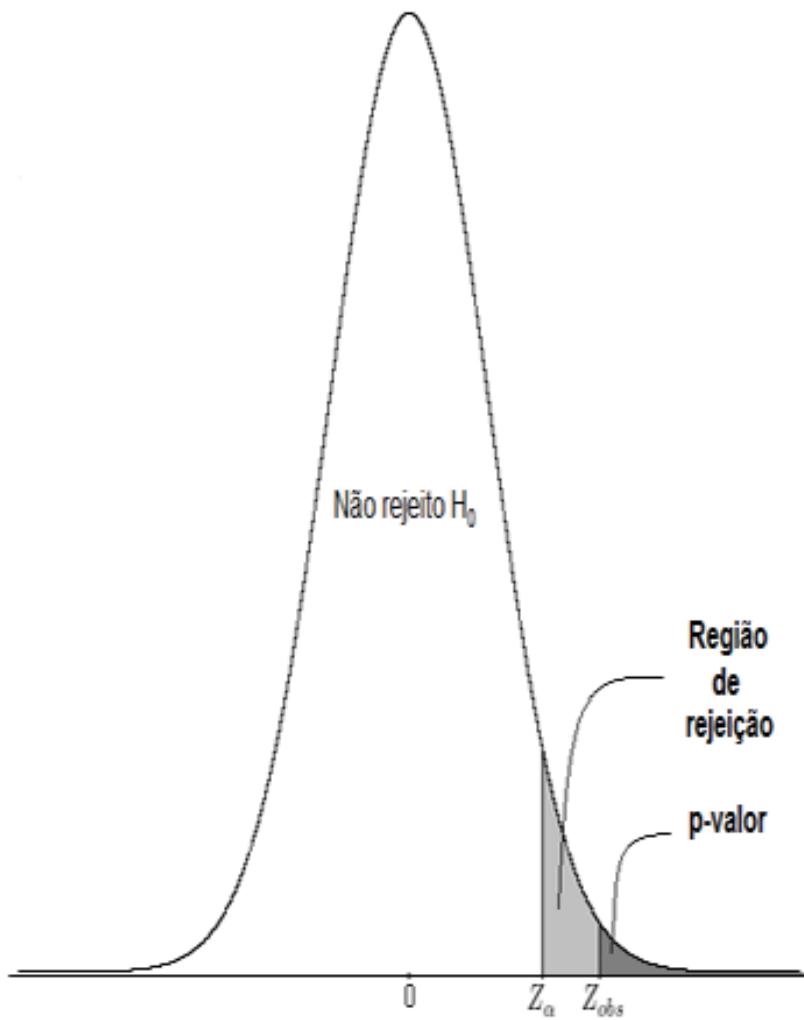
- É o menor nível de significância para o qual rejeitamos a hipótese nula.
- Um p-valor pequeno significa que a probabilidade de obter um valor da estatística de teste como o observado é muito improvável, consequentemente, rejeita-se a hipótese nula.

Teste bilateral

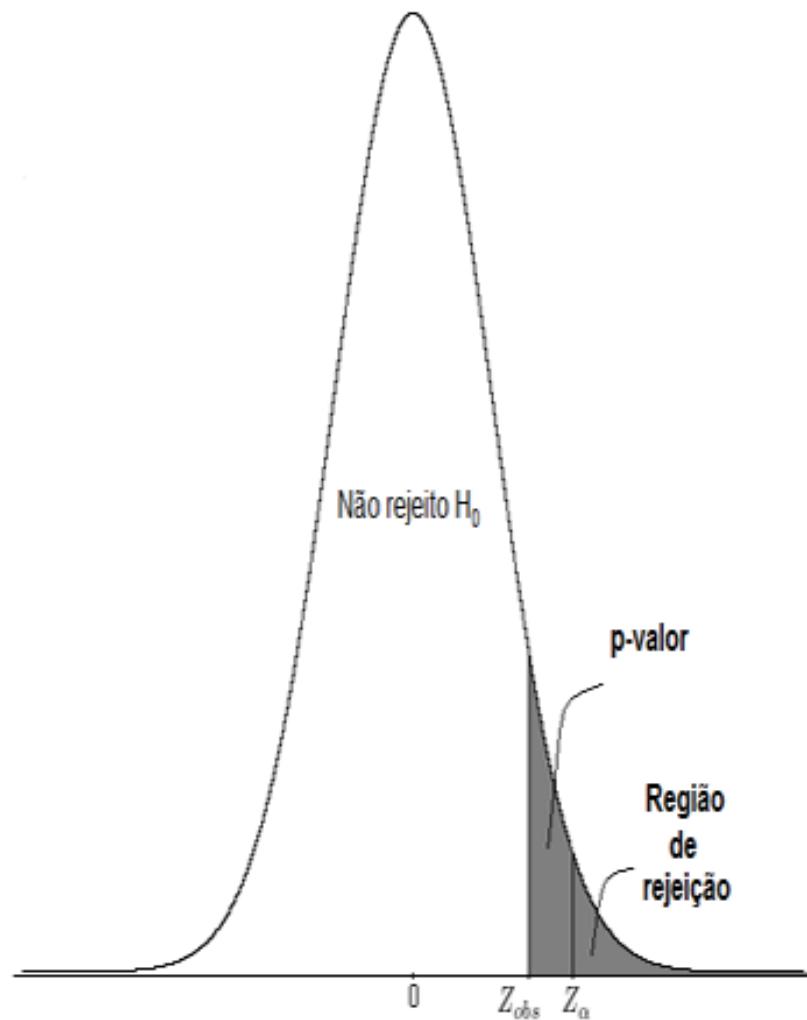


Teste unilateral

Teste unilateral à direita



Teste unilateral à direita



Observação: Se o p-valor é menor que o nível de significância proposto, então o valor da estatística Z_{obs} estará na região crítica e, portanto, rejeitamos a hipótese nula H_0 . Em contrapartida, se o p-valor é maior que o nível de significância, não rejeitamos a hipótese nula H_0 .

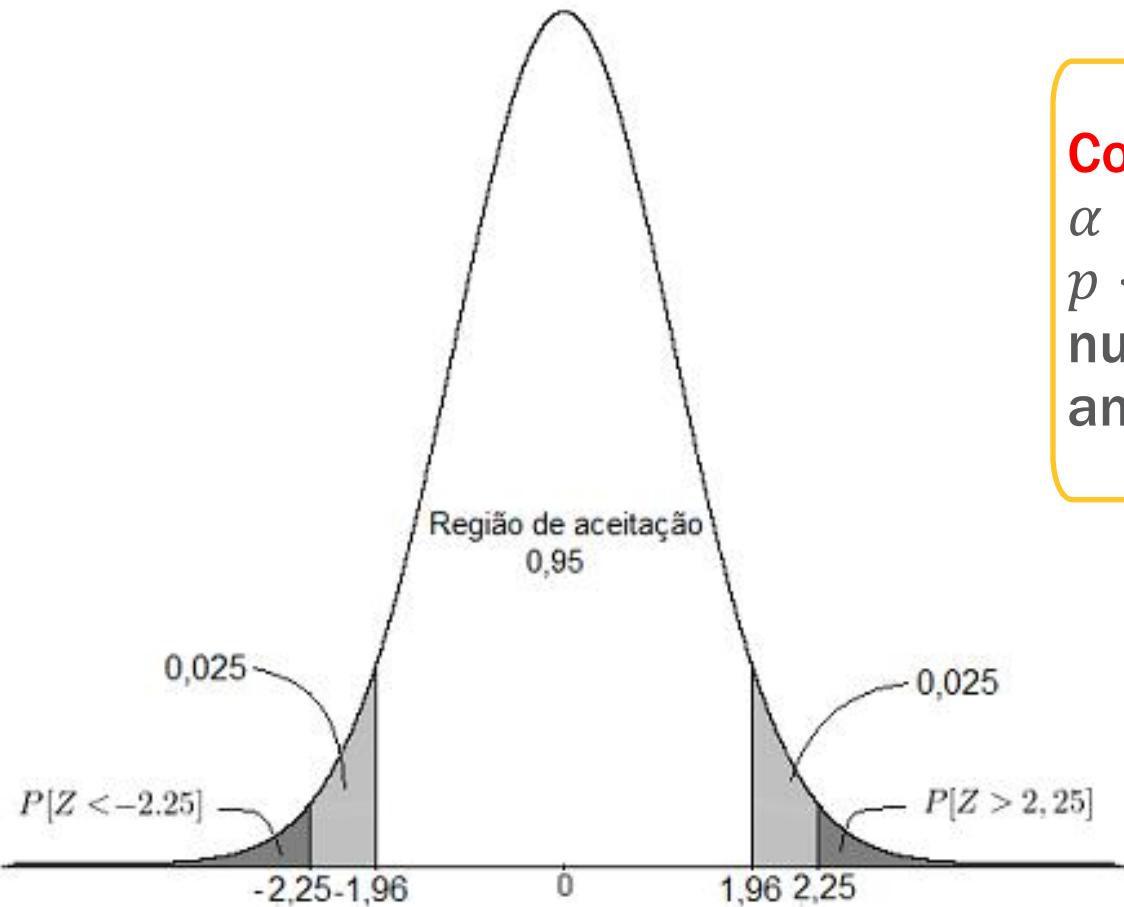
Exemplo 5.1.1.1: Um supervisor da qualidade quer testar, com base em uma amostra aleatória de tamanho $n = 35$ e, para um nível de significância $\alpha = 5\%$ se a profundidade média de um furo numa determinada peça é $72,4\text{ mm}$. O que podemos dizer se ele obteve $\bar{X} = 73,2\text{ mm}$ e se sabe, de informações anteriores, que $\sigma = 2,1\text{ mm}$?

Já vimos, neste exemplo, que o valor da estatística Z_{obs} é dado por:

$$Z_{obs} = 2,25$$

Desta forma, temos que $|Z_{obs}| = 2,25$ e, como o teste é bilateral, temos que o p-valor é dado por

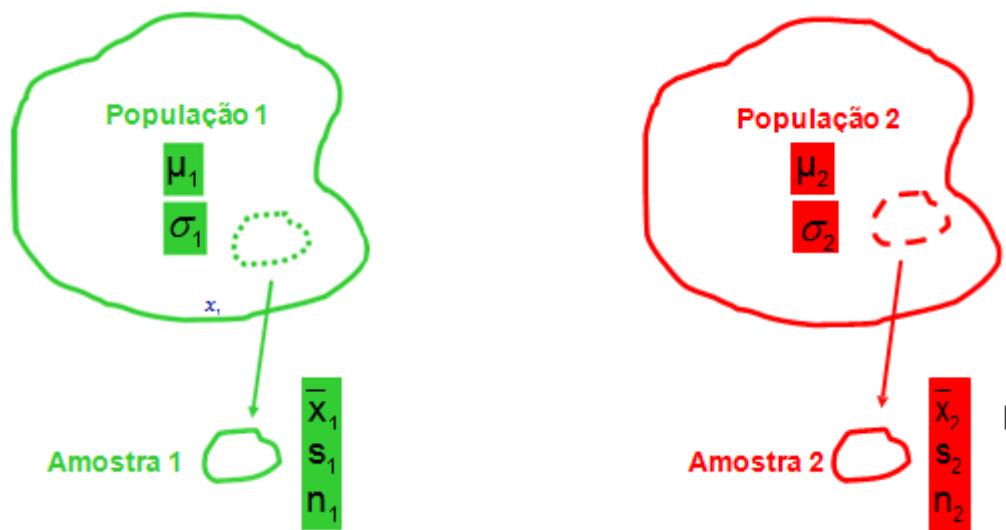
$$p = 2 \times P(Z > |Z_{obs}|) = 2 \times P(Z > 2,25) = 0,0244$$



Conclusão: Considerando $\alpha = 0,05$, como o p-valor $p < \alpha$, rejeitamos a hipótese nula. Como já sabíamos anteriormente.

TESTE “t”

- O teste t considera a comparação entre duas médias utilizando-se da variável t da distribuição t de Student. Se considerarmos duas amostras extraídas de duas populações 1 e 2 e quisermos verificar se essas populações são iguais ou diferentes baseados nos resultados das amostras selecionadas temos caracterizado um teste t.



TESTE “t”

- A aplicação do teste t nos fornecerá como resultado um valor p:
- valor $p \geq 0,05$ indica que não há evidências de que μ_1 seja diferente de μ_2
- valor $p < 0,05$ indica que há evidências suficientes para acreditar que
- μ_1 seja diferente de μ_2
- Cabe salientar que, para um nível de confiança estabelecido, o valor p encontrado depende:
 - da média das amostras (quanto mais diferentes menor o valor p)
 - do desvio padrão das amostras (quanto menores menor o valor p)
 - do tamanho das amostras (quanto maiores menor o valor p)

TESTE “t”

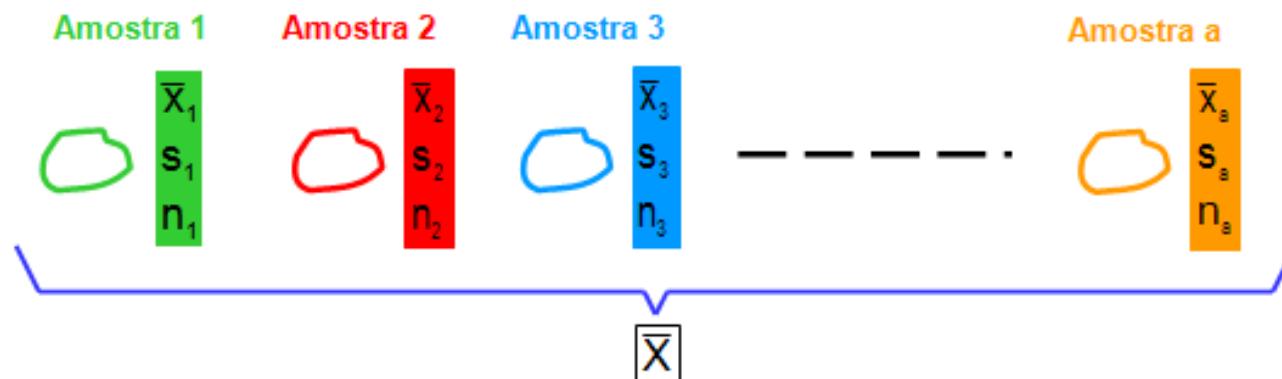
- Os testes t mais utilizados são de três tipos:

teste t para *variâncias equivalentes*: aplicado quando as amostras provem de populações que supostamente apresentam o mesmo nível de variância.

- teste t para *variâncias diferentes*: aplicado quando as amostras provem de populações que supostamente apresentam níveis de variância diferentes.
- teste t para *amostras pareadas*: utilizado nos casos em que o sujeito do teste pode interferir nos resultados

ANOVA

Considere uma quantidade “a” de amostras a serem comparadas, cada uma com sua média (\bar{x}), desvio padrão (s) e tamanho (n) e onde \bar{X} é a média de todas as amostras, conforme mostrado na figura abaixo:



$$S_{entre}^2 = \frac{n}{a-1} \sum_{i=1}^a (\bar{x}_i - \bar{X})^2$$

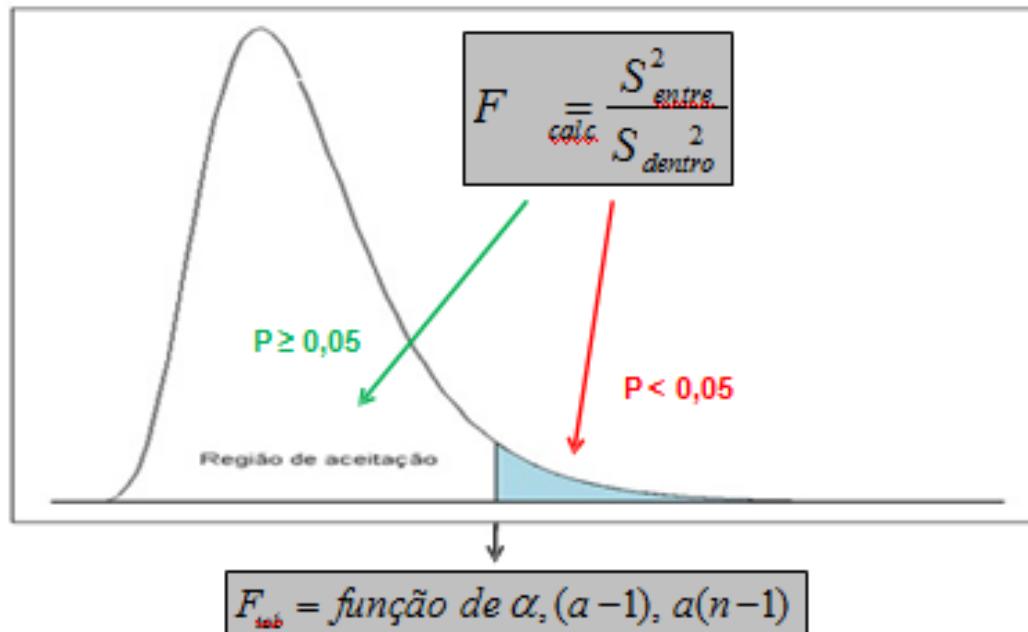
$$S_{dentro}^2 = \frac{\sum_{j=1}^a \sum_{i=1}^n (x_{ij} - \bar{x}_j)^2}{a(n-1)}$$

ANOVA

- $F_{\text{calc}} \leq F_{\text{tab}} \rightarrow \text{valor } p \geq 0,05$
- $F_{\text{calc}} > F_{\text{tab}} \rightarrow \text{valor } p < 0,05$

A figura a seguir ilustra esse procedimento:

ANOVA para nível de confiança de 95%:



Se valor $p \geq 0,05$ significa que não há evidências que uma ou mais das medias μ_i seja diferente das demais;

Se valor $p < 0,05$ significa que há evidências de que pelo menos uma das medias μ_i é diferente das demais.

Análise de Variância

ANOVA

Efeito fixo

ANOVA Um Fator

ANOVA

Ferramenta utilizada para analisar o comportamento de diversos tratamentos de um fator aplicados a um processo e/ou produto

Exemplo:

comportamento de uma característica da qualidade do processo em diversos níveis de um fator de controle.

Considere um fator A com k níveis, fixos. Esses dados, poder ser representados como na tabela a seguir.

Nível	Fator A	Soma	Média	Desvio Padrão
1	$y_{11} \ y_{12} \ \dots \ y_{1n_1}$	$y_{1.}$	$\bar{y}_{1.}$	$s_{1.}$
2	$y_{21} \ y_{22} \ \dots \ y_{2n_2}$	$y_{2.}$	$\bar{y}_{2.}$	$s_{2.}$
\vdots	$\vdots \ \vdots \ \vdots \ \vdots$	\vdots	\vdots	\vdots
K	$y_{k1} \ y_{k2} \ \dots \ y_{kn_k}$	$y_{k.}$	$\bar{y}_{k.}$	$s_{k.}$

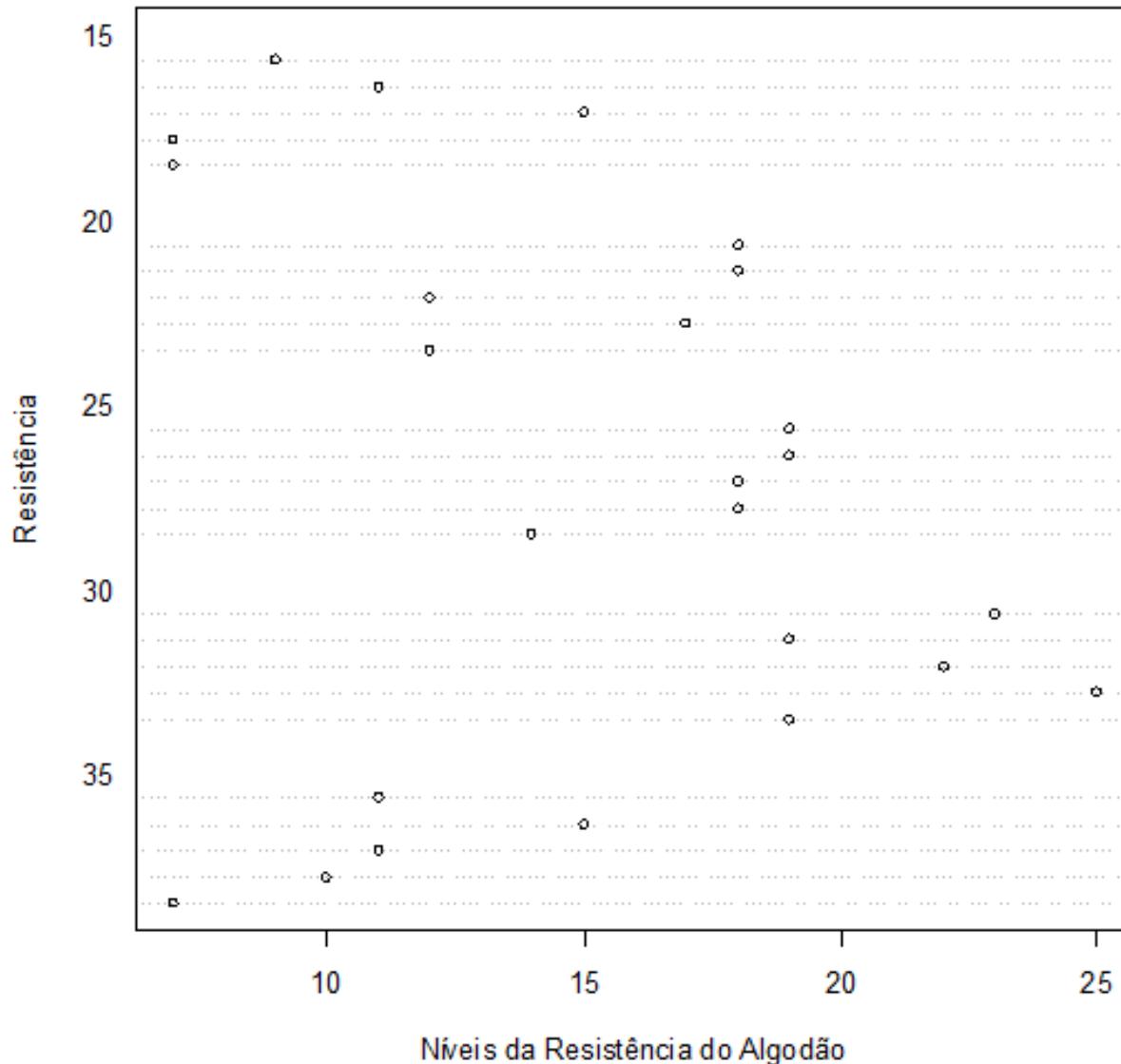
Exemplo

Considere o processo de produção de uma fibra sintética, onde o experimentador quer conhecer a influência da porcentagem de algodão na resistência da fibra. Para isto, foi realizado um experimento totalmente aleatorizado, onde diversos níveis de porcentagem de algodão foram avaliados com respeito à resistência da fibra. Um ponto importante na montagem do experimento é que para cada nível do fator (porcentagem de algodão), os outros fatores que influenciam o processo devem apresentar um padrão homogêneo de variabilidade. No experimento, tomamos 5 níveis para a porcentagem de algodão e 5 replicações.

Nível	Fator A					Soma	Média	Desvio Padrão
1	7	7	15	11	9	49	9,8	3,3
2	12	17	12	18	18	77	15,4	3,1
3	14	18	18	19	19	88	17,6	2,1
4	19	25	22	19	23	108	21,6	2,6
5	7	10	11	15	11	54	10,8	2,8
						$y_{..} = 376$	$\bar{y}_{..} = 15,04$	$s = 5,15$
								$s_{Entre} = 4,87$

Exemplo

Dotplot (Resistência da Fibra)



Modelo da ANOVA

O modelo de médias:

$$Y_{ij} = \mu_i + \varepsilon_{ij},$$

com, $j = 1, 2, \dots, n_i$
 $i = 1, 2, \dots, k$

Em que

- μ_i : média do nível i
- ε_{ij} : erro experimental

No modelo de efeito fixo, temos:

$$\mu = \frac{\sum_{i=1}^n n_i \mu_i}{N} = \frac{\sum_{i=1}^n n_i (\mu + \alpha_i)}{N} = \mu + \frac{\sum_{i=1}^n n_i \alpha_i}{N}$$



$$\sum_{i=1}^n n_i \alpha_i = 0$$

$$\alpha_i = \mu_i - \mu$$

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij},$$

com, $j = 1, 2, \dots, n_i$
 $i = 1, 2, \dots, k$

Em que

- μ : média geral
- α_i : efeito do nível i do fator
- ε_{ij} : erro experimental

Hipóteses

$\varepsilon_{ij} \sim N(0, \sigma^2)$ independentes e identicamente distribuídas.

$y_{ij} \sim N(\mu + \alpha_i, \sigma^2)$ independentes para $i = 1, \dots, k$.

Análise de variância - ANOVA

Teste de
Hipóteses



$$H_0: \mu_1 = \mu_2 = \dots = \mu_k$$

H_1 : pelo menos um é diferente.

No modelo de efeito fixo, temos:

$$\mu = \frac{\sum_{i=1}^n n_i \mu_i}{N} = \frac{\sum_{i=1}^n n_i (\mu + \alpha_i)}{N} = \mu + \frac{\sum_{i=1}^n n_i \alpha_i}{N}$$



$$\sum_{i=1}^n n_i \alpha_i = 0$$

Assim, podemos escrever as hipóteses, como:

$$H_0: \alpha_1 = \alpha_2 = \dots = \alpha_k = 0$$
$$H_1: \alpha_i \neq 0 \text{ (para algum } i = 1, \dots, k)$$



H_0 : Os níveis dos fatores são iguais
 H_1 : Pelo menos um nível do fator é diferente

Decomposição da Soma de Quadrados

A soma de quadrados total é definida como medida da variabilidade total dos dados



$$SQT = \sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{..})^2$$



$$\sum_{i=1}^k \sum_{j=1}^{n_i} [(y_{ij} - \bar{y}_{i.}) + (\bar{y}_{i.} - \bar{y}_{..})]^2$$



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

$$SQT = SQE + SQA$$

Notações:

- $y_{i.} = \sum_{j=1}^{n_i} y_{ij}$ =soma das observações do nível i do fator A
- $\bar{y}_{i.} = \frac{y_{i.}}{n_i}$ =média das observações do nível i do fator A
- $y_{..} = \sum_{i=1}^k \sum_{j=1}^{n_i} y_{ij}$ =soma de todas as observações
- $\bar{y}_{..} = \frac{y_{..}}{N}$ =média geral das observações

Sendo $N = \sum_{i=1}^k n_i$, total de observações

Decomposição da Soma de Quadrados

SQT

$$\sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{..})^2 = (N-1) \sum_{i=1}^k \sum_{j=1}^{n_i} \frac{(y_{ij} - \bar{y}_{..})^2}{N-1} = (N-1)s^2$$

$s^2 =$ variância dos dados

SQE

$$\sum_{i=1}^k \sum_{j=1}^{n_i} (n_i - 1) \frac{(y_{ij} - \bar{y}_{..})^2}{n_i - 1} = \sum_{i=1}^k (n_i - 1)s_i^2$$

$n_1 = \dots = n_k$

SQA

$$\sum_{i=1}^k \sum_{j=1}^{n_i} (\bar{y}_{i.} - \bar{y}_{..})^2 = \sum_{i=1}^k n_i (\bar{y}_{i.} - \bar{y}_{..})^2 = n(k-1) \sum_{i=1}^k \frac{(\bar{y}_{i.} - \bar{y}_{..})^2}{(k-1)} = n(k-1)s_{Entre}^2$$

Exemplo

Considere o processo de produção de uma fibra sintética, onde o experimentador quer conhecer a influência da porcentagem de algodão na resistência da fibra. Para isto, foi realizado um experimento totalmente aleatorizado, onde diversos níveis de porcentagem de algodão foram avaliados com respeito à resistência da fibra. Um ponto importante na montagem do experimento é que para cada nível do fator (porcentagem de algodão), os outros fatores que influenciam o processo devem apresentar um padrão homogêneo de variabilidade. No experimento, tomamos 5 níveis para a porcentagem de algodão e 5 replicações.

Nível	Fator A					Soma	Média	Desvio Padrão
1	7	7	15	11	9	49	9,8	3,3
2	12	17	12	18	18	77	15,4	3,1
3	14	18	18	19	19	88	17,6	2,1
4	19	25	22	19	23	108	21,6	2,6
5	7	10	11	15	11	54	10,8	2,8
						$y_{..} = 376$	$\bar{y}_{..} = 15,04$	$s = 5,15$
								$s_{Entre} = 4,87$

Exemplo

$$\bar{y}_{..} = 15,04$$

$$y_{1j} = 49, y_{2j} = 77, y_{3j} = 88, y_{4j} = 108 \text{ e } y_{5j} = 54$$

$$\bar{y}_{1.} = 9,8, \bar{y}_{2.} = 15,4, \bar{y}_{3.} = 17,6, \bar{y}_{4.} = 21,6 \text{ e } \bar{y}_{5.} = 10,8$$

$$s^2 = 26,54 \text{ e } S_{Entre}^2 = 23,79$$

SQT



$$(N - 1)s^2 = (25 - 1) * 26,54 = 636,96$$

SQE



$$\sum_{i=1}^{5} (n_i - 1)s_i^2 = (5 - 1) * [11,2 + 9,8 + 4,3 + 6,8 + 8,2] = 161,2$$

SQA



$$n(k - 1)s_{Entre}^2 = 5 * (5 - 1) * 23,79 = 475,76$$

Modelo da ANOVA: Graus de Liberdade e Estimativas da Variância

QME

$$E[QME] = E \left[\frac{SQE}{N - k} \right] = \frac{1}{N - k} E[SQE] = \sigma^2; \quad \text{e}$$

QMA

$$E[QMA] = E \left[\frac{SQA}{k - 1} \right] = \frac{1}{k - 1} E[SQA] = \sigma^2 + \frac{1}{k - 1} \sum_{i=1}^k n_i \alpha_i^2$$

Soma de Quadrados	Graus de Liberdade	Quadrados Médios
SQA	K-1	$\frac{SQA}{k - 1}$
SQE	N-k	$\frac{SQE}{N - k}$
SQT	N-1	

Modelo da ANOVA: Análise Estatística

Objetivo	Hipótese
efeito do fator A	$\begin{cases} H_0: \alpha_1 = \alpha_2 = \dots = \alpha_k = 0 \\ H_1: \alpha_i \neq 0 \text{ (para algum } i = 1, \dots, k) \end{cases}$

Sob H_0 :

Se $Y_{ij} \sim N(\mu, \sigma^2)$ i.i.d.

$$\frac{SQT}{\sigma^2} = \frac{SQE}{\sigma^2} + \frac{SQA}{\sigma^2}$$

Para estabelecer a independência entre as SQE e a SQA, utilizamos o teorema de Cochran.


$$\frac{SQT}{\sigma^2} = \sum_{i=1}^k \sum_{j=1}^{n_i} \left[\frac{y_{ij} - \bar{y}_{..}}{\sigma} \right]^2 \sim \chi_{N-1}^2$$
$$\frac{SQE}{\sigma^2} = \sum_{i=1}^k \frac{(n_i - 1)s_i^2}{\sigma^2} \sim \chi_{N-k}^2$$
$$\sum_{i=1}^k \frac{(n_i - 1)s_i^2}{\sigma^2} \sim \chi_{\sum_{i=1}^k (n_i - 1)}^2$$

Teste da ANOVA - Um Fator

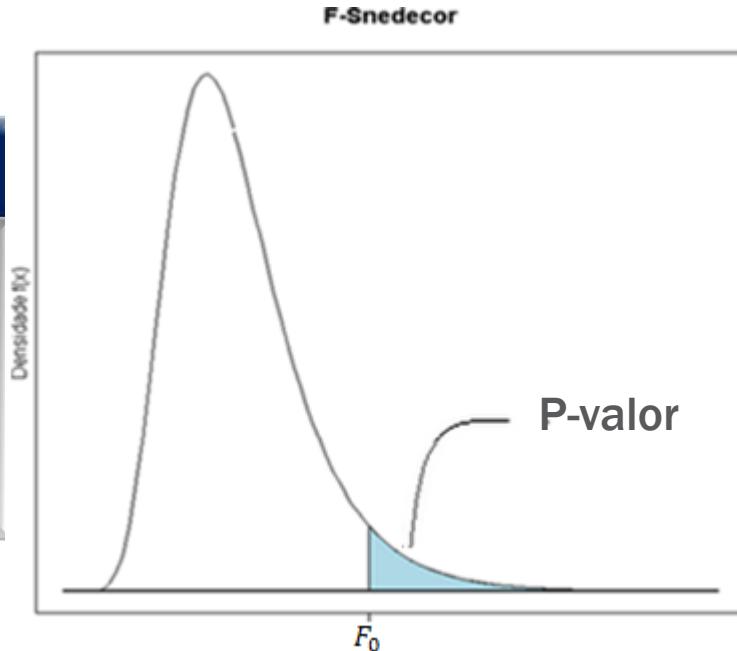
$\frac{SQA}{\sigma^2}$ têm distribuição Qui-Quadrado independentes

$\frac{SQE}{\sigma^2}$ têm distribuição Qui-Quadrado independentes



$$F_0 = \frac{\frac{SQA}{(k-1)}}{\frac{SQE}{(N-k)}} = \frac{QMA}{QME} \sim F_{(k-1; N-k)}$$

Fator	Soma de Quadrados	Graus de Liberdade	Quadrados Médios	F_0	P-valor
Fator	SQA	$k - 1$	$QMA = \frac{SQA}{k - 1}$		
Erro	SQE	$N - k$	$QME = \frac{SQE}{N - k}$	$F_0 = \frac{QMA}{QME}$	$P\text{-valor} = P[F_{(k-1, N-k)} > F_0 H_0]$
Total	SQT	$N - 1$			



$$P\text{-valor} = P[F_{(k-1, N-k)} > F_0 | H_0]$$

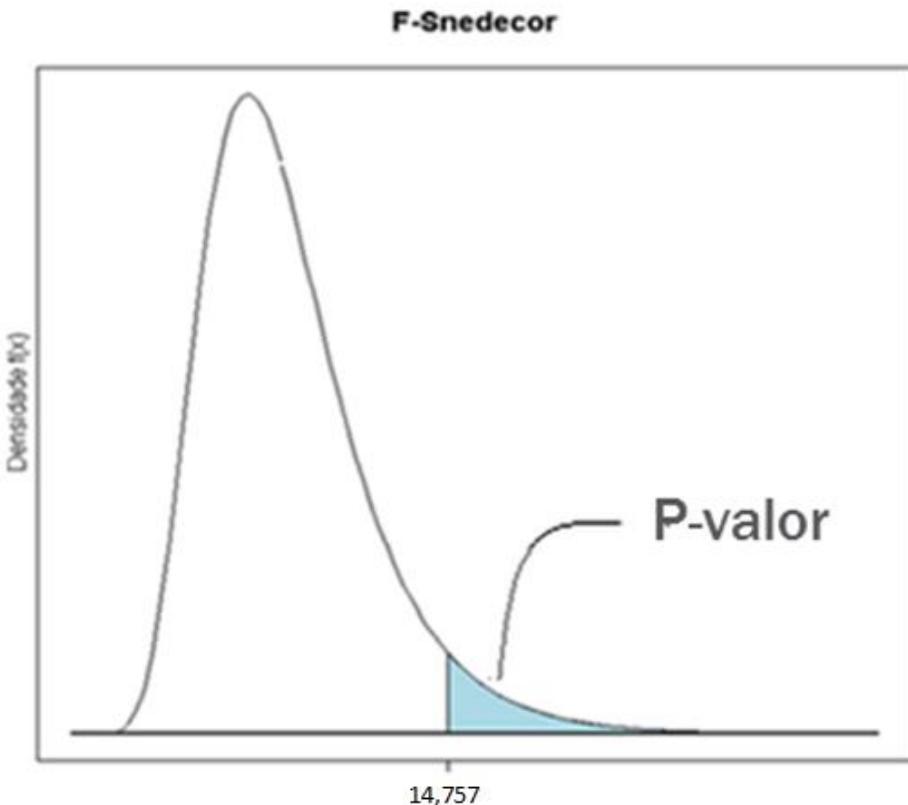
Exemplo

Considere o processo de produção de uma fibra sintética, onde o experimentador quer conhecer a influência da porcentagem de algodão na resistência da fibra. Para isto, foi realizado um experimento totalmente aleatorizado, onde diversos níveis de porcentagem de algodão foram avaliados com respeito à resistência da fibra. Um ponto importante na montagem do experimento é que para cada nível do fator (porcentagem de algodão), os outros fatores que influenciam o processo devem apresentar um padrão homogêneo de variabilidade. No experimento, tomamos 5 níveis para a porcentagem de algodão e 5 replicações.

Nível	Fator A					Soma	Média	Desvio Padrão
1	7	7	15	11	9	49	9,8	3,3
2	12	17	12	18	18	77	15,4	3,1
3	14	18	18	19	19	88	17,6	2,1
4	19	25	22	19	23	108	21,6	2,6
5	7	10	11	15	11	54	10,8	2,8
						$y_{..} = 376$	$\bar{y}_{..} = 15,04$	$s = 5,15$
								$s_{Entre} = 4,87$

Exemplo

Fator	Soma de Quadrados	Graus de Liberdade	Quadrados Médios	F_0	P-valor
Fator	475,76	$5 - 1 = 4$	$\frac{475,76}{4} = 118,94$		
Erro	161,20	$25 - 5 = 20$	$\frac{161,2}{20} = 8,06$	$F_0 = \frac{QMA}{QME} = 14,757$	P-valor = 0
Total	636,96	$25 - 1 = 24$			



Para $\alpha = 0,05$, obtemos que o p -valor = 0. Portanto, considerando 95% de confiança, rejeitamos H_0 , ou seja, pelo menos um nível do fator é diferente de zero.

Estimação dos Parâmetros do Modelo

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij},$$

com, $j = 1, 2, \dots, n_i$
 $i = 1, 2, \dots, k$

- $\hat{\mu} = \bar{y}_{..}$ estimador da média geral
- $\hat{\alpha}_i = \bar{y}_{i.} - \bar{y}_{..}$ estimador dos efeitos
- $\mu_i = \mu + \alpha_i$ a média do i-ésimo nível

Se assumirmos que os erros ε_{ij} são normalmente distribuídos, obtemos
 $\bar{y}_{i.} \sim N(\mu + \alpha_i, \frac{\sigma^2}{n_i})$



$$\frac{\frac{(\bar{y}_{i.} - \mu_i)}{\sqrt{\sigma^2/n_i}}}{\sqrt{QME/\sigma^2}} = \frac{(\bar{y}_{i.} - \mu_i)}{\sqrt{QME/n_i}} \sim t_{(N-k)}$$

Intervalo com confiança de $(1 - \alpha)100\%$ para μ_i é

$$\bar{y}_{i.} - t_{\left(\frac{\alpha}{2}, N-k\right)} \times \sqrt{\frac{QME}{n_i}} \leq \mu_i \leq \bar{y}_{i.} + t_{\left(\frac{\alpha}{2}, N-k\right)} \times \sqrt{\frac{QME}{n_i}}$$

Estimação dos Parâmetros do Modelo

Vamos construir o Intervalo com confiança de $(1 - \alpha)100\%$ para a diferença entre as médias de dois níveis, para isto,

$$\frac{\bar{y}_{i.} - \mu_i}{\sqrt{\sigma^2/n_i}} \sim N(0,1)$$

$$\frac{\bar{y}_{i.} - \bar{y}_{l.} - (\mu_i - \mu_l)}{\sigma \sqrt{\frac{1}{n_i} + \frac{1}{n_l}}} \sim N(0,1)$$

com, $i = 1, 2, \dots, k$ e
são independentes

$$\frac{\bar{y}_{i.} - \bar{y}_{l.} - (\mu_i - \mu_l)}{\sigma \sqrt{\frac{1}{n_i} + \frac{1}{n_l}}} = \frac{\bar{y}_{i.} - \bar{y}_{l.} - (\mu_i - \mu_l)}{\sqrt{QME \left(\frac{1}{n_i} + \frac{1}{n_l} \right)}} \sim t_{(N-k)}$$

Portanto, o Intervalo com confiança de $(1 - \alpha)100\%$ para a diferença entre as médias de dois níveis é

$$(\bar{y}_{i.} - \bar{y}_{l.}) - \Delta \leq \mu_i - \mu_l \leq (\bar{y}_{i.} - \bar{y}_{l.}) + \Delta \quad \text{em que } \Delta = t_{(1-\frac{\alpha}{2}, N-k)} \times \sqrt{QME \left(\frac{1}{n_i} + \frac{1}{n_l} \right)}$$

Exemplo

Considere o processo de produção de uma fibra sintética, onde o experimentador quer conhecer a influência da porcentagem de algodão na resistência da fibra. Para isto, foi realizado um experimento totalmente aleatorizado, onde diversos níveis de porcentagem de algodão foram avaliados com respeito à resistência da fibra. Um ponto importante na montagem do experimento é que para cada nível do fator (porcentagem de algodão), os outros fatores que influenciam o processo devem apresentar um padrão homogêneo de variabilidade. No experimento, tomamos 5 níveis para a porcentagem de algodão e 5 replicações.

Nível	Fator A					Soma	Média	Desvio Padrão
1	7	7	15	11	9	49	9,8	3,3
2	12	17	12	18	18	77	15,4	3,1
3	14	18	18	19	19	88	17,6	2,1
4	19	25	22	19	23	108	21,6	2,6
5	7	10	11	15	11	54	10,8	2,8
						$y_{..} = 376$	$\bar{y}_{..} = 15,04$	$s = 5,15$
								$s_{Entre} = 4,87$

Exemplo

$$\bar{y}_i = 21,6$$

$$t_{\left(\frac{\alpha}{2}, N-k\right)} = 2,086$$

$$QME = 8,06$$

$$n_i = 5$$

Intervalo com confiança de 95% para a média do nível 4 (30% de algodão na fibra) é calculado da seguinte forma:

$$21,6 - 2,086 \times \sqrt{\frac{8,06}{5}} \leq \mu_4 \leq 21,6 + 2,086 \times \sqrt{\frac{8,06}{5}}$$

Com isso, obtemos $18,95 \leq \mu_4 \leq 24,25$

Um intervalo com confiança de 95% para a diferença entre a média dos níveis 4 e 5 (30% e 35% de algodão na fibra) é dado por

$$(24,6 - 10,8) - 3,74 \leq \mu_4 - \mu_5 \leq (24,6 - 10,8) + 3,74 \Rightarrow \\ \Rightarrow 7,05 \leq \mu_4 - \mu_5 \leq 14,55$$

$$\text{Em que } \Delta = t_{\left(1-\frac{\alpha}{2}; N-k\right)} \times \sqrt{QME \left(\frac{1}{n_i} + \frac{1}{n_l} \right)} = 2,086 \sqrt{8,06 \left(\frac{1}{5} + \frac{1}{5} \right)} = 3,74$$

Exemplo

ANOVA

DADOS DO PROCESSO

Tabela da Anova					
	G.L.	Soma Quad	Quadrado Médio	Estat. F	P-valor
Fator	4	475,76	118,94	14,76	0
Resíduos	20	161,2	8,06		

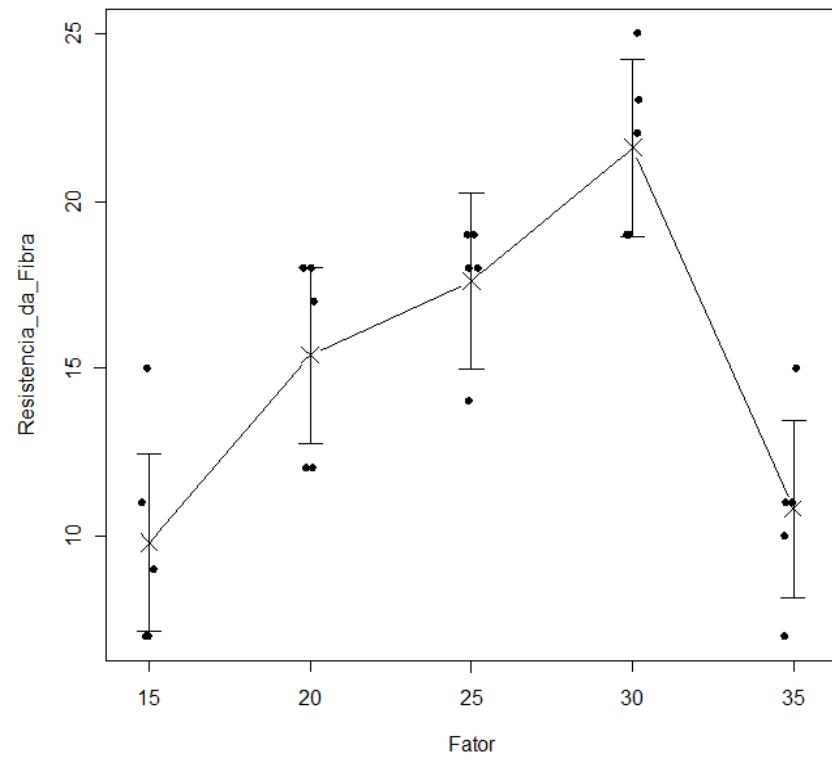
Intervalo de Confiança das Médias

Fator	Média	Desvio Padrão	Limite Inferior	Limite Superior
15	9,8	2,59	7,15	12,45
20	15,4	2,59	12,75	18,05
25	17,6	2,59	14,95	20,25
30	21,6	2,59	18,95	24,25
35	10,8	2,59	8,15	13,45

Intervalo de confiança dos Efeitos

Fator	Limite Inferior	Efeito	Limite Superior
15	7,15	9,8	12,45
20	12,75	15,4	18,05
25	14,95	17,6	20,25
30	18,95	21,6	24,25
35	8,15	10,8	13,45

Intervalo de Confiança das Médias de Intervalos



Análise de Resíduos

Hipóteses

- ◻ $\varepsilon_{ij} \sim N(0, \sigma^2)$ independentes e identicamente distribuídos;
- ◻ Modelo para os dados:

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

Na prática, precisamos verificar se estas suposições são válidas. Violações nestas suposições são verificadas através dos resíduos

O resíduo para a j-ésima observação do nível i

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

Em que \hat{y}_{ij} é uma estimativa da observação y_{ij}

$$\hat{y}_{ij} = \bar{y}_{..} + (\bar{y}_{i..} - \bar{y}_{..}) = \hat{\mu} + \hat{\alpha}_i = \bar{y}_{i..}$$

Se o modelo for adequado, os resíduos observados devem refletir essas propriedades.

Propriedades dos resíduos

Média



$$\bar{e} = \frac{\sum_{i=1}^n e_i}{n} = 0$$

Variância



$$\frac{\sum_{i=1}^n (e_i - \bar{e})^2}{n - 2} = \frac{\sum_{i=1}^n e_i^2}{n - 2} = \frac{SQE}{n - 2} = QME$$

Exemplo

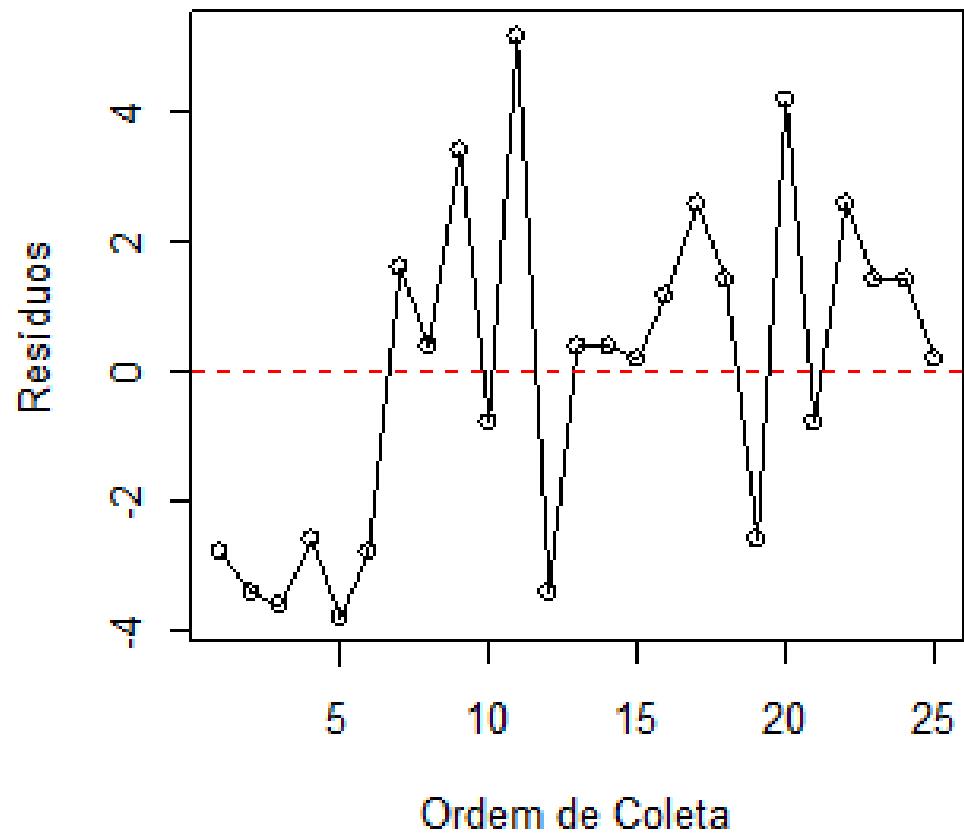
Considere o processo de produção de uma fibra sintética, onde o experimentador quer conhecer a influência da porcentagem de algodão na resistência da fibra. Para isto, foi realizado um experimento totalmente aleatorizado, onde diversos níveis de porcentagem de algodão foram avaliados com respeito à resistência da fibra. Um ponto importante na montagem do experimento é que para cada nível do fator (porcentagem de algodão), os outros fatores que influenciam o processo devem apresentar um padrão homogêneo de variabilidade. No experimento, tomamos 5 níveis para a porcentagem de algodão e 5 replicações.

Nível	Fator A					Soma	Média	Desvio Padrão
1	7	7	15	11	9	49	9,8	3,3
2	12	17	12	18	18	77	15,4	3,1
3	14	18	18	19	19	88	17,6	2,1
4	19	25	22	19	23	108	21,6	2,6
5	7	10	11	15	11	54	10,8	2,8
						$y_{..} = 376$	$\bar{y}_{..} = 15,04$	$s = 5,15$
								$s_{Entre} = 4,87$

Exemplo

Algodão	Resíduos					$\hat{y}_j = \bar{y}_L$
15	-2,8	-2,8	1,2	-0,8		9,8
20	-3,4	1,6	2,6	2,6		15,4
25	-3,6	0,4	1,4	1,4		17,6
30	-2,6	3,4	-2,6	1,4		21,6
35	-3,8	-0,8	4,2	0,2		10,8

Resíduos x Ordem de Coleta



Análise de Resíduos

Grande parte dos problemas que encontramos na prática, são solucionados, considerando algumas suposições iniciais



Assumir a normalidade dos dados é o primeiro passo que tomamos para simplificar sua análise. Para dar suporte a esta suposição, consideramos os testes:

- Teste de Anderson-Darling;
- Teste Kolmogorov-Smirnov;
- Teste Shapiro-Wilk.

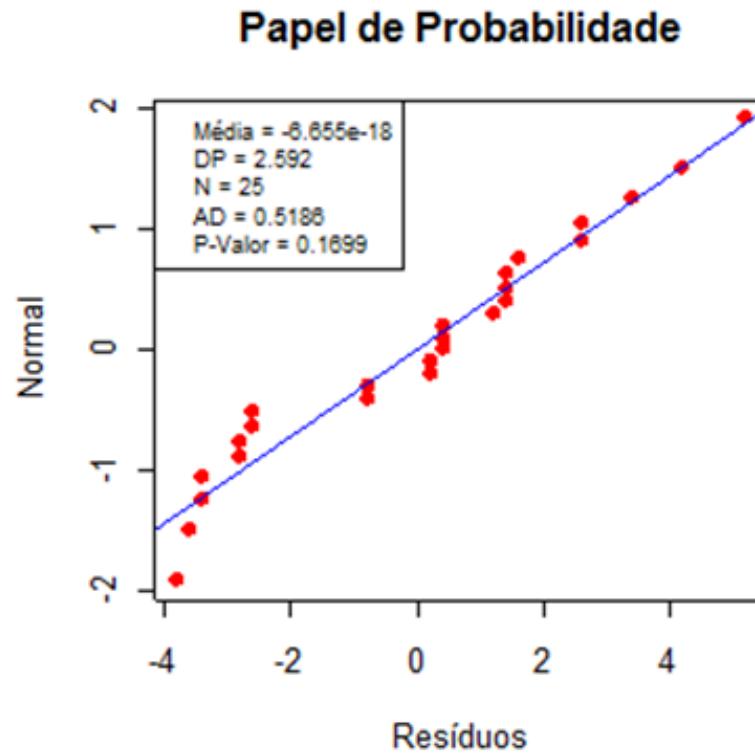
E o gráfico “Papel de Probabilidade”

Papel de probabilidade e Teste de Anderson-Darling

Avaliamos a normalidade dos resíduos através do gráfico “papel de probabilidade” e do teste de Anderson-Darling. Consideramos a seguinte hipótese:

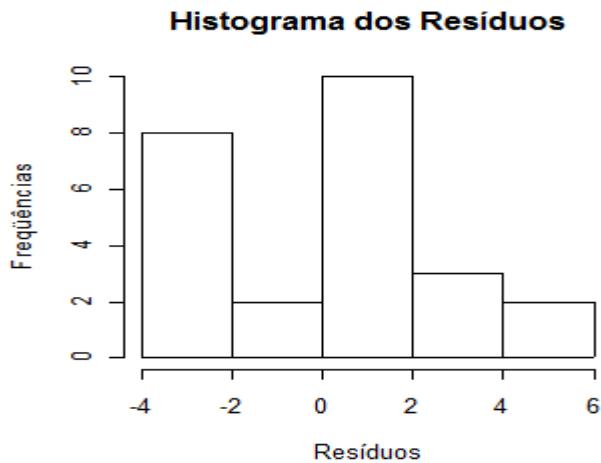
$$\begin{cases} H_0: \text{os resíduos seguem uma distribuição Normal} \\ H_1: \text{os resíduos Não seguem uma distribuição Normal} \end{cases}$$

Para o exemplo, como o P-valor é alto (aproximadamente 0,16) não rejeitamos a hipótese de normalidade dos resíduos.



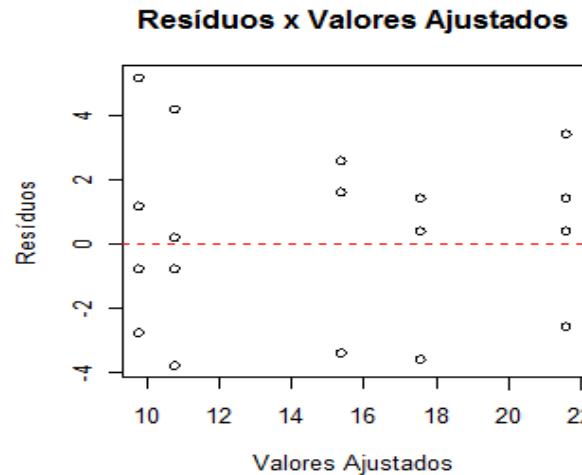
Papel de probabilidade e Teste de Anderson-Darling

Também desenvolvemos o histograma dos resíduos para avaliar sua dispersão e distribuição.

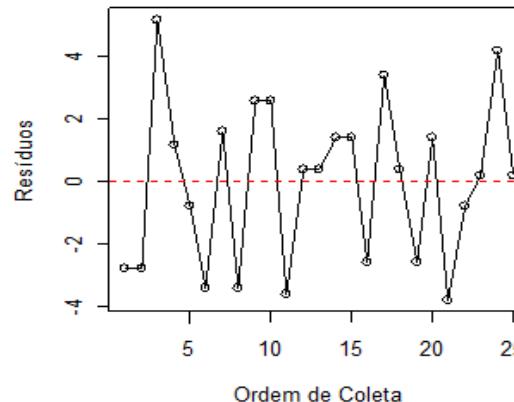


Resíduos versus valores ajustados

Com esse gráfico temos indícios sobre o comportamento da variância dos resíduos com relação aos valores ajustados.



Resíduos x Ordem de Coleta



Resíduos versus a ordem de coleta dos dados

Com esse gráfico obtemos indícios da independência ou não entre os resíduos. Se algum comportamento sistemático for observado no gráfico, temos indícios de que alguma variável "extra" influenciou nos resultados do experimento, fato que viola uma das premissas básicas da ANOVA e compromete nossas conclusões.

Teste de Igualdade das Variâncias

No modelo da ANOVA 1 fator, não necessariamente temos que as variâncias sejam iguais, ou seja, podemos ter variâncias diferentes (modelo heterocedástico). Considere as hipóteses:

$$\begin{cases} H_0: \sigma_1^2 = \sigma_2^2 = \dots = \sigma_k^2 \\ H_1: \text{pelo menos 1 dos } \sigma_i^2 \text{ são diferentes para } i = 1, 2, \dots, k. \end{cases}$$

Teste de Bartlet



Utilizado quando a hipótese de normalidade dos resíduos não é rejeitada.

Teste de Levene



Utilizado quando a hipótese de normalidade dos resíduos é rejeitada.

Teste de Bartlet

A estatística do teste proposta por Bartlett é dada por

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



Em que

- $q = (N - k) \ln s_p^2 - \sum_{i=1}^k [(n_i - 1) \ln s_i^2]$
- $c = 1 + \frac{1}{3(k-1)} \left(\sum_{i=1}^k \frac{1}{n_i - 1} - \frac{1}{N-k} \right)$
- $s_p^2 = \sum_{i=1}^k \frac{(n_i - 1)s_i^2}{N-k}$ e $s_i^2 = \sum_{j=1}^{n_i} \frac{(y_{ij} - \bar{y}_{i.})^2}{n_i - 1}$

Rejeitamos H_0 quando

$$B_0 > Q_{[1-\alpha;k-1]}$$

onde $Q_{[1-\alpha;k-1]}$ representa o percentil com $(1 - \alpha)100\%$ da distribuição Qui-quadrado com $(k-1)$ graus de liberdade.

O P-valor é calculado por

$$\text{P-valor} = P[\chi_{(k-1)}^2 > B_0 | H_0]$$

Exemplo

Considere o processo de produção de uma fibra sintética, onde o experimentador quer conhecer a influência da porcentagem de algodão na resistência da fibra. Para isto, foi realizado um experimento totalmente aleatorizado, onde diversos níveis de porcentagem de algodão foram avaliados com respeito à resistência da fibra. Um ponto importante na montagem do experimento é que para cada nível do fator (porcentagem de algodão), os outros fatores que influenciam o processo devem apresentar um padrão homogêneo de variabilidade. No experimento, tomamos 5 níveis para a porcentagem de algodão e 5 replicações.

Nível	Fator A					Soma	Média	Desvio Padrão
1	7	7	15	11	9	49	9,8	3,3
2	12	17	12	18	18	77	15,4	3,1
3	14	18	18	19	19	88	17,6	2,1
4	19	25	22	19	23	108	21,6	2,6
5	7	10	11	15	11	54	10,8	2,8
						$y_{..} = 376$	$\bar{y}_{..} = 15,04$	$s = 5,15$
								$s_{Entre} = 4,87$

Exemplo

TESTES DE VARIÂNCIA

DADOS DO PROCESSO

Informação	Valor
Bartlett (estatística do teste)	0,93
Graus de Liberdade	4
P-valor	0,92

Fator	Intervalos de Confiança para o Desvio-padrão		
	Limite Inferior	Desvio Padrão	Limite Superior
15	1,736	3,347	14,712
20	1,624	3,130	13,762
25	1,076	2,074	9,116
30	1,353	2,608	11,463
35	1,486	2,864	12,588

A estatística do teste proposta por Bartlett é dada por

$$B_0 = \frac{q}{c} = \frac{1,0264}{1,1} = 0,93$$

Em que

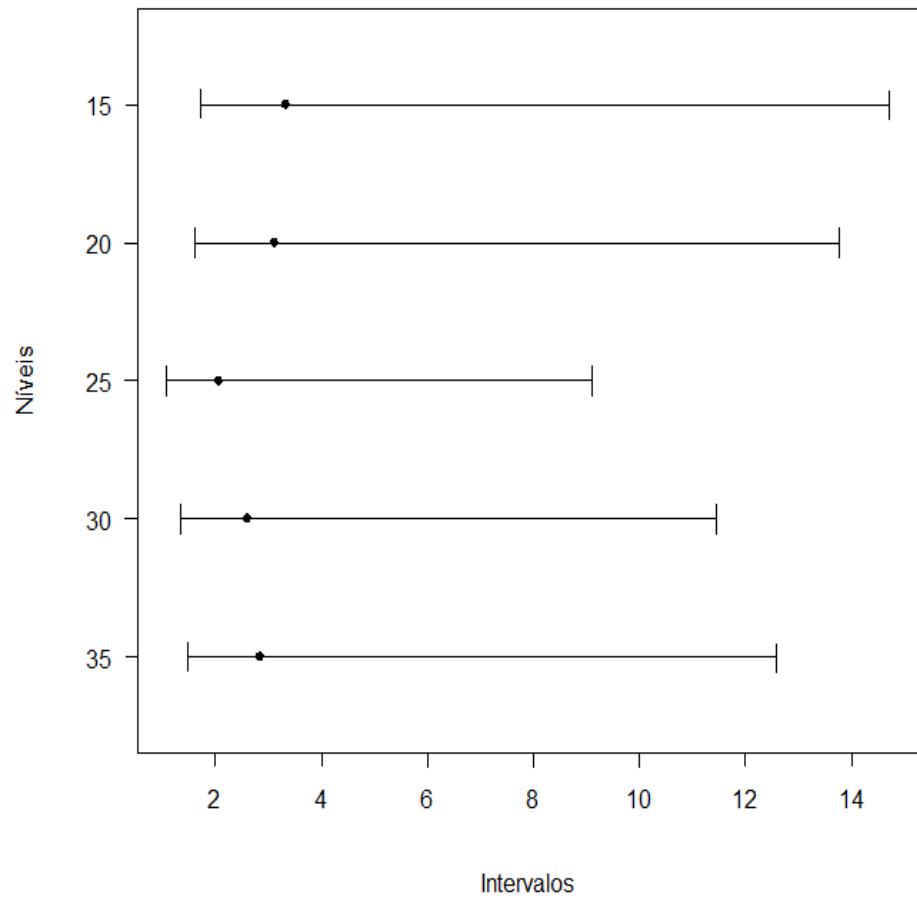
$q = (N - k) \ln s_p^2 - \sum_{i=1}^k [(n_i - 1) \ln s_i^2] = 1,0264$

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

O P-valor é calculado por

$$\text{P-valor} = P[\chi_{(k-1)}^2 > B_0 | H_0] = P[\chi_{(k-1)}^2 > 0,93 | H_0] = 0,92$$

Intervalos de Confiança para o Desvio-padrão



Teste de Levene

Levene (1960) propôs a seguinte transformação:

$$z_{ij} = |x_{ij} - \bar{x}_{i\cdot}| \text{ ou } z_{ij} = |x_{ij} - \tilde{x}_{i\cdot}|$$

com, $i = 1, 2, \dots, k$
 $j = 1, 2, \dots, n_i$



Em que

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

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

Exemplo

Considere o processo de produção de uma fibra sintética, onde o experimentador quer conhecer a influência da porcentagem de algodão na resistência da fibra. Para isto, foi realizado um experimento totalmente aleatorizado, onde diversos níveis de porcentagem de algodão foram avaliados com respeito à resistência da fibra. Um ponto importante na montagem do experimento é que para cada nível do fator (porcentagem de algodão), os outros fatores que influenciam o processo devem apresentar um padrão homogêneo de variabilidade. No experimento, tomamos 5 níveis para a porcentagem de algodão e 5 replicações.

Nível	Fator A					Soma	Média	Desvio Padrão
1	7	7	15	11	9	49	9,8	3,3
2	12	17	12	18	18	77	15,4	3,1
3	14	18	18	19	19	88	17,6	2,1
4	19	25	22	19	23	108	21,6	2,6
5	7	10	11	15	11	54	10,8	2,8
						$y_{..} = 376$	$\bar{y}_{..} = 15,04$	$s = 5,15$
								$s_{Entre} = 4,87$

SUMÁRIO BÁSICO E POR FATOR

DADOS DO PROCESSO

Fator	Mediana
15	9
20	17
25	18
30	22
35	11

k	Algodão%	Resistência da fibra				
1	15	2	2	6	2	0
2	20	5	0	5	1	1
3	25	4	0	0	1	1
4	30	3	3	0	3	1
5	35	4	1	0	4	0

$$z_{11} = |x_{ij} - \tilde{x}_i| = |7 - 9| = 2$$

Exemplo

TESTES DE VARIÂNCIA

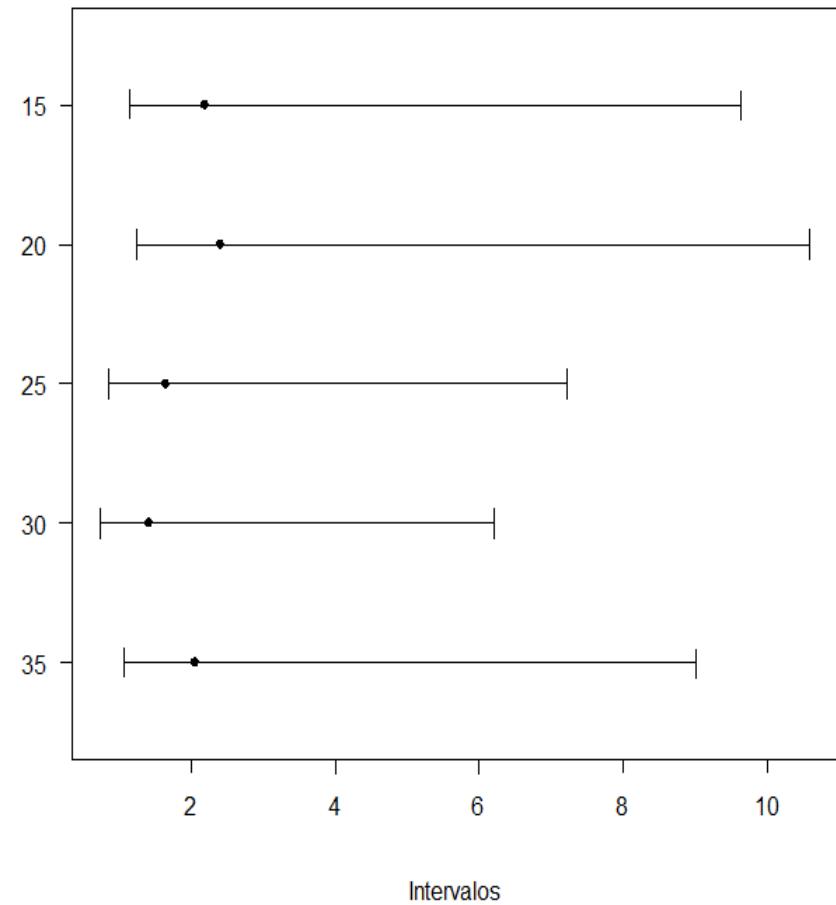
DADOS DO PROCESSO

Informação	Valor
Levene (estatística do teste)	0,26
Graus de Liberdade	4
P-valor	0,90

Fator	Intervalos de Confiança para o Desvio-padrão		
	Limite Inferior	Desvio Padrão	Limite Superior
15	1,137	2,191	9,631
20	1,249	2,408	10,587
25	0,853	1,643	7,223
30	0,734	1,414	6,217
35	1,063	2,049	9,009

Conclusão: Como o p-valor é maior que 5%, não temos evidências para rejeitar a hipótese de igualdade de variâncias.

Intervalos de Confiança para o Desvio-padrão



Teste de Comparações Múltiplas

Quando os resultados da Análise de Variância (ANOVA) levam à rejeição da hipótese nula, ($H_0 = \mu_1 = \dots = \mu_k$), que representa a afirmação de que todas as médias (tratamentos) são iguais, temos evidências de que as médias entre os níveis diferem significativamente.



Os testes de comparações múltiplas permitem identificar essas diferenças entre pares de médias específicos ou em combinações lineares das médias.

Definição



Família: é um conjunto de inferências para o qual é importante levar em conta alguma medida de erro global.



Erro do tipo I: rejeitar a hipótese nula, quando esta é verdadeira.



Erro do tipo II: aceitar a hipótese nula, quando esta é falsa .

Teste de Comparações Múltiplas

Abordagem por comparação

Uma possível abordagem para o problema de comparação múltipla é fazer cada comparação independentemente usando um procedimento estatístico adequado. Por exemplo, um teste de hipótese estatístico pode ser usado para comparar cada par de médias, μ_i e μ_j

Dificuldade com a abordagem
“por comparação” para
comparações múltiplas



- possibilidade do aumento da probabilidade do Erro Tipo I ou (equivalentemente)
- possibilidade de diminuição do nível de confiança global.

Exemplo: consideremos dois testes de hipóteses independentes cada um ao nível de significância α . Assim, a probabilidade que nenhum tenha Erro Tipo I é $(1 - \alpha)^2$. Em outras palavras, a probabilidade de ao menos um Erro do Tipo I é $1 - (1 - \alpha)^2$.

Geralmente, para testar as diferenças entre cada par k de médias é necessário o total de $c = \frac{1}{2}k(k - 1)$ testes t ao nível de significância α .

Teste de Comparações Múltiplas

Abordagem por comparação

A chance de encontrarmos ao menos uma diferença incorreta é
 $\alpha_c = 1 - (1 - \alpha)^c$.

Para $k \geq 3$, além de termos α_c maior que α , temos ainda que α_c se aproxima de 1 conforme k aumenta.

A tabela a seguir ilustra tal situação, em que calculamos a probabilidade de ao menos uma rejeição incorreta da hipótese nula para diferentes valores de c .

c	%	c	%	c	%
1	5	10	40,12	15	53,67
2	9,75	11	43,12	20	64,15
3	14,26	12	45,96	30	78,53
4	18,55	13	48,67	40	87,14
5	22,62	14	51,23	50	92,3

Teste de Comparações Múltiplas

Taxa de erros

Tipos de Taxas de Erro

- ❑ Quando uma família é composta por vários testes de hipóteses e cada teste de hipótese é realizado ao mesmo nível de significância α , então α é a taxa de erro por comparação (TPC) (per-comparison error rate), isto é, a probabilidade de rejeitarmos incorretamente cada uma das hipóteses nulas que compõem a família.
- ❑ Uma taxa de erro mais apropriada é chamada de taxa de erro da família dos testes (familywise error rate (FWER)), que é a probabilidade de rejeitarmos incorretamente ao menos umas das hipóteses nulas que compõem a família.
- ❑ Há ainda um terceiro tipo de taxa de erro conhecido como taxa de erro por família (TPF)(per family error rate), que não é uma probabilidade como as outras taxas são, mas representa o valor esperado dos erros na família. Por exemplo, assumimos que a hipótese nula global é verdadeira, se cada dos c testes é realizado com probabilidade de Erro Tipo I α/c , o valor esperado do Erro Tipo I é $c \times (\alpha/c) = \alpha$.

$$TPC \leq FWER \leq TPF$$

Teste de Comparações Múltiplas

Métodos de Comparações Múltiplas (MCMs)

- ❑ são procedimentos estatísticos designados para ter em conta e controlar o aumento da probabilidade global do Erro do Tipo I ou a diminuição do intervalo de confiança global.

Classificações (MCMs)

“Etapa única” (Single Step)

- ❑ Teste de Tukey (dados balanceados);
- ❑ Teste de Tukey-kramer (dados não balanceados) ;
- ❑ Teste de Dunnet;
- ❑ Teste de Scheffé;
- ❑ Teste de HSU;
- ❑ Teste de Bonferroni.

“Etapa abaixo” (Step-down)

- ❑ Teste de Fisher;
- ❑ Procedimento de Holm.

“Etapa acima” (Step-up)

- ❑ Procedimento de Hochberg;
- ❑ Procedimento de Hommel.

Teste de Comparações Múltiplas

Teste de Tukey

- ❑ O Teste proposto por Tukey (1953) é também conhecido como teste de Tukey da diferença honestamente significativa (*honestly significant difference*)(HSD) e teste de Tukey da diferença totalmente significativa (*wholly significant difference*)(WSD).
- ❑ É um teste exato em que, para a família de todas as $c = \frac{k(k-1)}{2}$ comparações duas a duas, a taxa de erro da família dos testes (FWER) é exatamente α (e o intervalo de confiança é exatamente $1 - \alpha$).

A estratégia de Tukey consiste em definir a menor diferença significativa. Tal procedimento utiliza a amplitude da distribuição studentizada.

Suponhamos que temos k observações independentes, Y_1, \dots, Y_k , de uma distribuição normal com média μ e variância σ^2 . Seja w a amplitude para esse conjunto de observações

$$w = \max(Y_i) - \min(Y_i)$$

Teste de Comparações Múltiplas

Teste de Tukey

Suponhamos que temos uma estimativa s^2 da variância σ^2 , que é baseada nos $N - k$ graus de liberdade e é independente de Y_i , em que N é o número total de observações.

$$\frac{w}{s} = q(k ; N - k)$$

é chamada *amplitude studentizada*

Tamanhos de amostras iguais
(dados balanceados)



- O teste de Tukey declara duas médias significativamente diferentes se o valor absoluto de suas diferenças amostrais ultrapassar

$$TSD = q_\alpha(k ; N - k) \sqrt{\frac{QME}{n}}$$



- Um intervalo de confiança de $100(1 - \alpha)\%$ para a diferença entre todos os pares das médias é dado como

$$\bar{y}_{i.} - \bar{y}_{j.} - q_\alpha(k ; N - k) \sqrt{\frac{QME}{n}} \leq \mu_i - \mu_j \leq \bar{y}_{i.} - \bar{y}_{j.} + q_\alpha(k ; N - k) \sqrt{\frac{QME}{n}}, \quad i \neq j$$

Tamanhos de amostras diferentes
(dados não balanceados)



- O teste de Tukey declara duas médias significativamente diferentes se o valor absoluto de suas diferenças amostrais ultrapassar

$$TSD = q_\alpha(k ; N - k) \sqrt{QME \left(\frac{1}{n_i} + \frac{1}{n_j} \right)}$$



- Um intervalo de confiança de $100(1 - \alpha)\%$ para a diferença entre todos os pares das médias é dado como

$$\bar{y}_{i.} - \bar{y}_{j.} - q_\alpha(k ; N - k) \sqrt{QME \left(\frac{1}{n_i} + \frac{1}{n_j} \right)} \leq \mu_i - \mu_j \leq \bar{y}_{i.} - \bar{y}_{j.} + q_\alpha(k ; N - k) \sqrt{QME \left(\frac{1}{n_i} + \frac{1}{n_j} \right)}, \quad i \neq j$$

Exemplo

Considere o processo de produção de uma fibra sintética, onde o experimentador quer conhecer a influência da porcentagem de algodão na resistência da fibra. Para isto, foi realizado um experimento totalmente aleatorizado, onde diversos níveis de porcentagem de algodão foram avaliados com respeito à resistência da fibra. Um ponto importante na montagem do experimento é que para cada nível do fator (porcentagem de algodão), os outros fatores que influenciam o processo devem apresentar um padrão homogêneo de variabilidade. No experimento, tomamos 5 níveis para a porcentagem de algodão e 5 replicações.

Nível	Fator A					Soma	Média	Desvio Padrão
1	7	7	15	11	9	49	9,8	3,3
2	12	17	12	18	18	77	15,4	3,1
3	14	18	18	19	19	88	17,6	2,1
4	19	25	22	19	23	108	21,6	2,6
5	7	10	11	15	11	54	10,8	2,8
						$y_{..} = 376$	$\bar{y}_{..} = 15,04$	$s = 5,15$
								$s_{Entre} = 4,87$

Exemplo

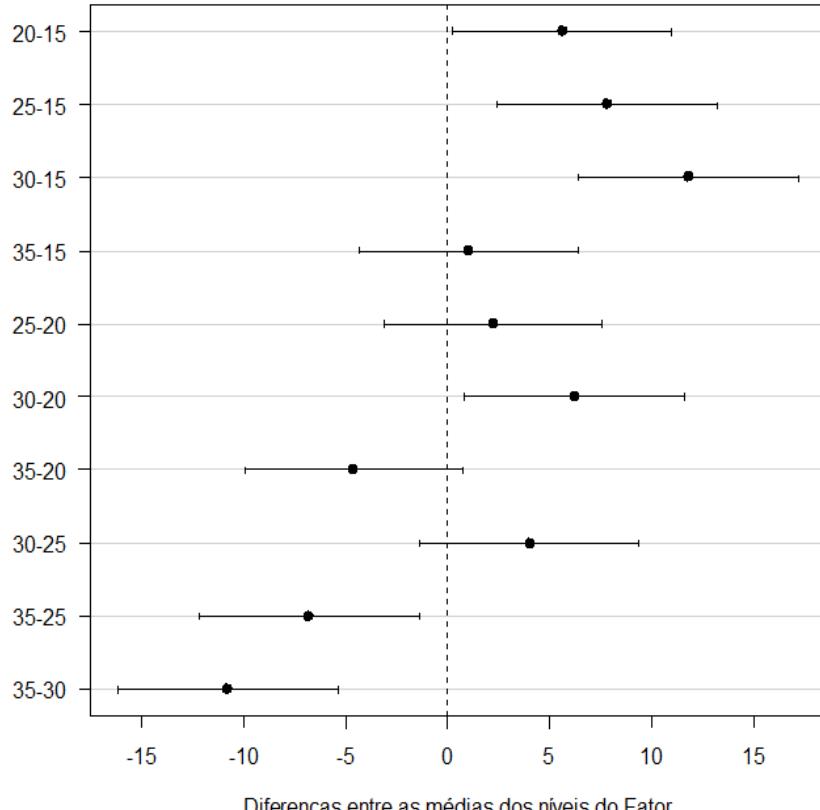
TESTE DE TUKEY

DADOS DO PROCESSO

Tabela da Anova

	G.L.	Soma de Quadrados	Quadrado Médio	Estat. F	P-valor
Fator	4	475,76	118,94	14,76	9,13E-06
Resíduos	20	161,2	8,06		

Intervalos de Confiança (95%)



Como os dados são balanceados, temos que:

$$TSD = q_{\alpha}(k ; N - k) \sqrt{\frac{QME}{n}}$$

$$= q_{0,05}(5 ; 20) \sqrt{\frac{8,06}{5}} = 5,373$$

Rejeitamos a igualdade entre dois níveis se:

$$|\bar{y}_{i\cdot} - \bar{y}_{j\cdot}| > 5,373$$

Níveis	Centro	Limite.Inferior	Limite.Superior	P-valor
20-15	5,6	0,23	10,97	0,039
25-15	7,8	2,43	13,17	0,0026
30-15	11,8	6,43	17,17	0,0000190
35-15	1	-4,37	6,37	0,98
25-20	2,2	-3,17	7,57	0,74
30-20	6,2	0,83	11,57	0,019
35-20	-4,6	-9,97	0,77	0,12
30-25	4	-1,37	9,37	0,21
35-25	-6,8	-12,17	-1,43	0,0091
35-30	-10,8	-16,17	-5,43	0,000062

Teste de Comparações Múltiplas

Teste de Fisher
Teste da diferença mínima significativa (least significant difference (LSD) test)

- ❑ O método de Fisher para comparar todos pares de médias controla a taxa de erro ao nível de significância α para cada comparação dois a dois, mas não controla a taxa de erro do experimento

- ❑ Esse procedimento usa a estatística t para testar $H_0: \mu_i = \mu_j$, em que

$$t_0 = \frac{\bar{y}_{i\cdot} - \bar{y}_{j\cdot}}{\sqrt{QME\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}}$$

- ❑ O procedimento de Fisher consiste em realizar testes t múltiplos, cada um ao nível de significância α , somente se o teste preliminar é significante ao nível α .
- ❑ Este pode ser visto como um procedimento de duas etapas em que a hipótese nula H_0 é testada no primeiro passo por um teste F de nível α .
- ❑ Se o teste F não é significativo, o procedimento termina sem precisar fazer inferências detalhadas nas diferenças dos pares das médias; caso contrário, cada diferença de par é testada por um teste t com nível α de significância.

Teste de Comparações Múltiplas

Teste de Fisher

O LSD controla a taxa de erro do experimento ao nível α sobre H_0 devido a “proteção” fornecida para essa hipótese pelo teste F preliminar. No entanto, em outras configurações (hipóteses) de médias verdadeiras, a taxa de erro do experimento pode ser maior que α .

Tamanhos de amostras iguais
(dados balanceados)



- O teste de Fisher considera duas médias significativamente diferentes se o valor absoluto de suas diferenças amostrais ultrapassar

$$LSD = t_{(\frac{\alpha}{2}, N-k)} \sqrt{2 \times \frac{QME}{n}}$$

Tamanhos de amostras diferentes
(dados não balanceados)



- O teste de Fisher considera duas médias significativamente diferentes se o valor absoluto de suas diferenças amostrais ultrapassar

$$LSD = t_{(\frac{\alpha}{2}, N-k)} \sqrt{QME \left(\frac{1}{n_i} + \frac{1}{n_j} \right)}$$

Exemplo

Considere o processo de produção de uma fibra sintética, onde o experimentador quer conhecer a influência da porcentagem de algodão na resistência da fibra. Para isto, foi realizado um experimento totalmente aleatorizado, onde diversos níveis de porcentagem de algodão foram avaliados com respeito à resistência da fibra. Um ponto importante na montagem do experimento é que para cada nível do fator (porcentagem de algodão), os outros fatores que influenciam o processo devem apresentar um padrão homogêneo de variabilidade. No experimento, tomamos 5 níveis para a porcentagem de algodão e 5 replicações.

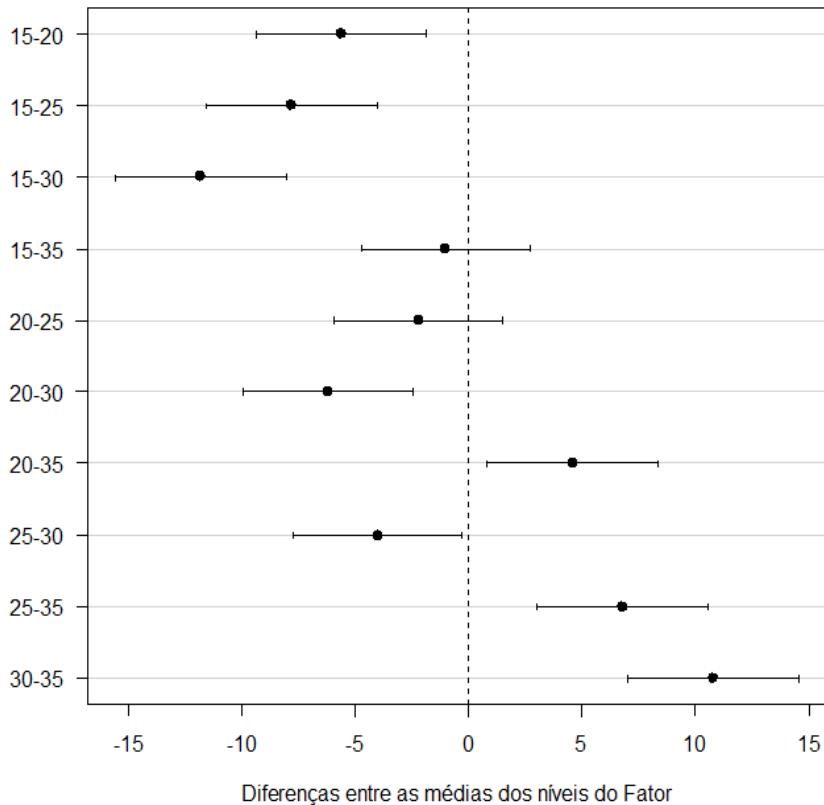
Nível	Fator A					Soma	Média	Desvio Padrão
1	7	7	15	11	9	49	9,8	3,3
2	12	17	12	18	18	77	15,4	3,1
3	14	18	18	19	19	88	17,6	2,1
4	19	25	22	19	23	108	21,6	2,6
5	7	10	11	15	11	54	10,8	2,8
						$y_{..} = 376$	$\bar{y}_{..} = 15,04$	$s = 5,15$
								$s_{Entre} = 4,87$

Exemplo

TESTE DE FISHER

DADOS DO PROCESSO

Intervalos de confiança (95%)



Como os dados são balanceados, temos que:

$$LSD = t_{(\frac{\alpha}{2}, N-k)} \sqrt{2 \times \frac{QME}{n}}$$
$$= t_{(0,025; 25-5)} \sqrt{2 \times \frac{8,06}{5}} = 3,7455$$

Rejeitamos a igualdade entre dois níveis se:

$$|\bar{y}_{i\cdot} - \bar{y}_{j\cdot}| > 3,7455$$

Diferença entre Níveis	Média	LI	LS	P-Valor
15-20	-5,6	-9,35	-1,85	0,0054
15-25	-7,8	-11,55	-4,05	0,0003
15-30	-11,8	-15,55	-8,05	0
15-35	-1	-4,75	2,75	0,58
20-25	-2,2	-5,95	1,55	0,23
20-30	-6,2	-9,95	-2,45	0,0025
20-35	4,6	0,85	8,35	0,019
25-30	-4	-7,75	-0,25	0,038
25-35	6,8	3,05	10,55	0,0012
30-35	10,8	7,05	14,55	0

Teste de Comparações Múltiplas

Teste ou Procedimento de Bonferroni

- ❑ Consiste na realização de um teste t para cada par de médias a uma taxa de erro por comparação (TPC) de $\frac{\alpha}{\binom{k}{2}}$. Usando esse teste, o nível de significância da família é no máximo α , para qualquer configuração (formação) das médias da população
- ❑ O teste de Bonferroni protege a taxa de erro da família dos testes.
- ❑ O teste de Bonferroni pode ser usado para quaisquer que sejam os dados balanceados ou não balanceados. Não é um teste exato, sendo baseado em uma aproximação conhecida como primeira desigualdade de Bonferroni.
- ❑ Em algumas situações, o teste de Bonferroni se mostra bastante “conservativo” (fraco), isto é, a taxa de erro da família de testes (FWER) é muito menor do que o nível de significância α estabelecido.

Teste de Comparações Múltiplas

Teste ou Procedimento de Bonferroni

Para a família de todas as comparações duas a duas, o procedimento de Bonferroni irá produzir intervalos de confiança maiores que o teste de Tukey ou Tukey-Kramer.

Tamanhos de amostras iguais
(dados balanceados)



- O teste de Bonferroni considera duas médias significativamente diferentes se o valor absoluto de suas diferenças amostrais ultrapassar

$$LSD = t_{(\alpha^*, N-k)} \sqrt{2 \times \frac{QME}{n}} \quad \alpha^* = \frac{\alpha}{2c}$$

Tamanhos de amostras diferentes
(dados não balanceados)



- O teste de Bonferroni considera duas médias significativamente diferentes se o valor absoluto de suas diferenças amostrais ultrapassar

$$LSD = t_{(\alpha^*, N-k)} \sqrt{QME \left(\frac{1}{n_i} + \frac{1}{n_j} \right)}$$

Exemplo

Considere o processo de produção de uma fibra sintética, onde o experimentador quer conhecer a influência da porcentagem de algodão na resistência da fibra. Para isto, foi realizado um experimento totalmente aleatorizado, onde diversos níveis de porcentagem de algodão foram avaliados com respeito à resistência da fibra. Um ponto importante na montagem do experimento é que para cada nível do fator (porcentagem de algodão), os outros fatores que influenciam o processo devem apresentar um padrão homogêneo de variabilidade. No experimento, tomamos 5 níveis para a porcentagem de algodão e 5 replicações.

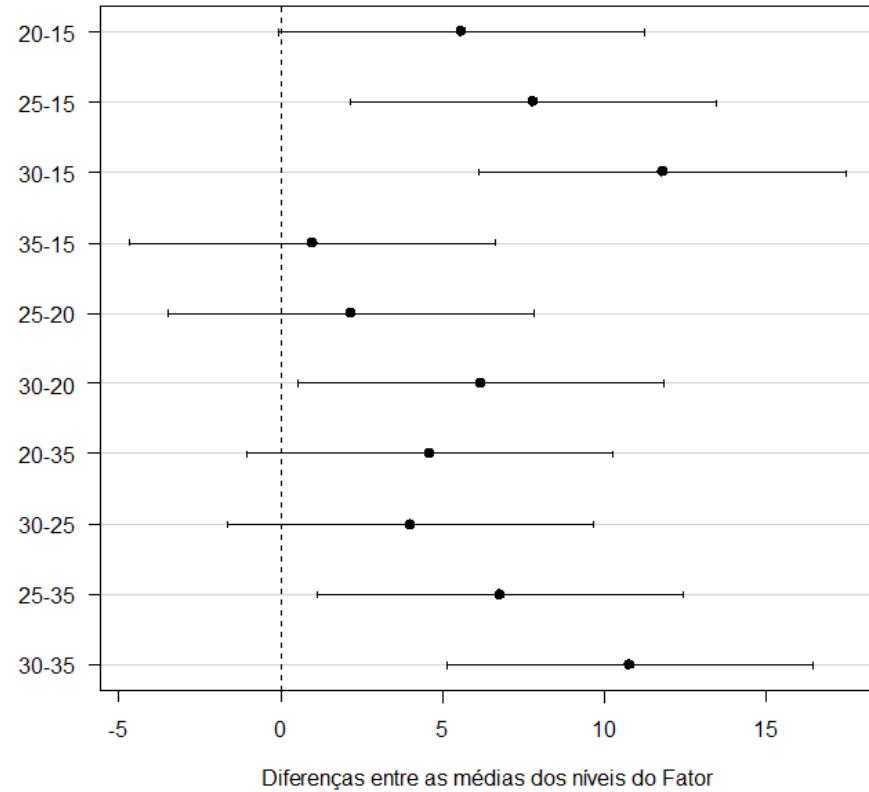
Nível	Fator A					Soma	Média	Desvio Padrão
1	7	7	15	11	9	49	9,8	3,3
2	12	17	12	18	18	77	15,4	3,1
3	14	18	18	19	19	88	17,6	2,1
4	19	25	22	19	23	108	21,6	2,6
5	7	10	11	15	11	54	10,8	2,8
						$y_{..} = 376$	$\bar{y}_{..} = 15,04$	$s = 5,15$
								$s_{Entre} = 4,87$

Exemplo

TESTE DE FISHER-BONFERRONI

DADOS DO PROCESSO

Intervalos de Confiança (95%)



Como os dados são balanceados, temos que:

$$LSD = t_{(\frac{\alpha}{2 \times c}, N-k)} \sqrt{2 \times \frac{QME}{n}}$$
$$= t_{(0,05; 25-5)} \sqrt{2 \times \frac{8,06}{5}} = 5,662$$

Rejeitamos a igualdade entre dois níveis se:

$$|\bar{y}_{i \cdot} - \bar{y}_{j \cdot}| > 5,662$$

Diferença entre Níveis	Média	LI	LS	P-Valor
20-15	5,6	-0,06	11,26	0,054
25-15	7,8	2,14	13,46	0,0031
30-15	11,8	6,14	17,46	0,000021
35-15	1	-4,66	6,66	1
25-20	2,2	-3,46	7,86	1
30-20	6,2	0,54	11,86	0,025
20-35	4,6	-1,06	10,26	0,186
30-25	4	-1,66	9,66	0,38
25-35	6,8	1,14	12,46	0,012
30-35	10,8	5,14	16,46	0,00007

Teste de Comparações Múltiplas

Teste de Scheffé

- ❑ O método proposto por Scheffé (1959) é também conhecido como teste de Scheffé da diferença completamente significativa (fully significant difference (FSD)) e como teste de Scheffé da diferença globalmente significativa (globally significant difference(GSD)).
- ❑ É um método exato no sentido em que, para as famílias (finitas) envolvendo todos os contrastes das k médias, a FWER é exatamente α .
- ❑ O Teste de Scheffé pode ser usado quando as comparações são selecionadas depois de olhar para os dados e incluem os contrastes, que nem todos são aos pares.
- ❑ Também pode ser utilizado quando um grande número de contrastes, nem todos aos pares, são especificados antes de coletar os dados.

Teste de Comparações Múltiplas

Teste de Scheffé

O método de Scheffé também pode ser usado para a família de todas as comparações duas a duas, mas quase sempre resultará em intervalos de confiança maiores que os métodos estudados anteriormente (Tukey, Tukey-Kramer, Fisher e Bonferroni).

Teste de Scheffé



- O Teste de Scheffé considera duas médias significativamente diferentes se o valor absoluto de suas diferenças amostrais ultrapassar

$$FSD = \sqrt{(k - 1)F_{(\alpha, k-1, N-k)}} \sqrt{QME \left(\frac{1}{n_i} + \frac{1}{n_j} \right)}$$

Intervalo de confiança para o contraste



$$\sum_{i=1}^k c_i \bar{y}_i \pm \sqrt{(k - 1)F_{(\alpha, k-1, N-k)}} \sqrt{QME \left(\frac{1}{n_i} + \frac{1}{n_j} \right)}$$

Intervalo de confiança para $\mu_i - \mu_j$



$$\bar{y}_{i\cdot} - \bar{y}_{j\cdot} \pm FSD$$

Exemplo

Considere o processo de produção de uma fibra sintética, onde o experimentador quer conhecer a influência da porcentagem de algodão na resistência da fibra. Para isto, foi realizado um experimento totalmente aleatorizado, onde diversos níveis de porcentagem de algodão foram avaliados com respeito à resistência da fibra. Um ponto importante na montagem do experimento é que para cada nível do fator (porcentagem de algodão), os outros fatores que influenciam o processo devem apresentar um padrão homogêneo de variabilidade. No experimento, tomamos 5 níveis para a porcentagem de algodão e 5 replicações.

Nível	Fator A					Soma	Média	Desvio Padrão
1	7	7	15	11	9	49	9,8	3,3
2	12	17	12	18	18	77	15,4	3,1
3	14	18	18	19	19	88	17,6	2,1
4	19	25	22	19	23	108	21,6	2,6
5	7	10	11	15	11	54	10,8	2,8
						$y_{..} = 376$	$\bar{y}_{..} = 15,04$	$s = 5,15$
								$s_{Entre} = 4,87$

TESTE DE SCHEFFE

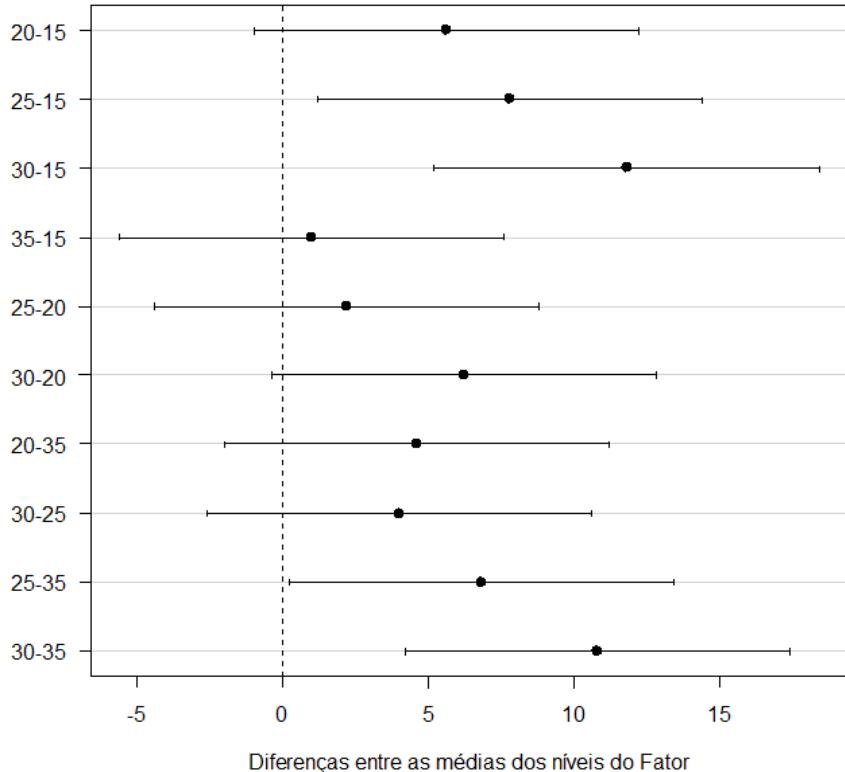
DADOS DO PROCESSO

Exemplo

Tabela da Anova

G.L.	Soma de Quadrados	Quadrado Médio	Estat. F	P-valor
Fator	4	118,94	14,76	9,128E-06
Resíduos	20	161,2	8,06	

Intervalos de Confiança (95%)



$$FSD = \sqrt{(k-1)F_{(\alpha, k-1, N-k)}} \sqrt{QME \left(\frac{1}{n_i} + \frac{1}{n_j} \right)} = \\ = \sqrt{(5-1)F_{(0,05,5-1,25-4)}} \sqrt{8,06 \left(\frac{1}{5} + \frac{1}{5} \right)} = 6,079$$

Rejeitamos a igualdade entre dois níveis se:

$$|\bar{y}_{i\cdot} - \bar{y}_{j\cdot}| > 6,079$$

Níveis	Centro	Limite.Inferior	Limite.Superior	P-valor
20-15	5,6	-1,01	12,21	0,081
25-15	7,8	1,19	14,41	0,0076
30-15	11,8	5,19	18,41	0,000079
35-15	1	-5,61	7,61	0,99
25-20	2,2	-4,41	8,81	0,82
30-20	6,2	-0,41	12,81	0,044
20-35	4,6	-2,01	11,21	0,20
30-25	4	-2,61	10,61	0,33
25-35	6,8	0,19	13,41	0,023
30-35	10,8	4,19	17,41	0,00024

Teste de Comparações Múltiplas

Teste de Dunnett

Comparações de médias com um controle ou com um valor referência

- ❑ Dunnert (1955) foi pioneiro no conceito de que, quando um controle está presente, as comparações de interesse preliminar podem ser as comparações de cada novo tratamento com o controle.

- ❑ Por exemplo, o controle pode ser um placebo, um tratamento “padrão”, ou qualquer outro tratamento específico (como uma nova droga).

- ❑ Suponhamos que μ_1, \dots, μ_{j-1} são as médias dos novos tratamentos e μ_j é a média do controle. Quando realizamos comparações múltiplas com um controle, os parâmetros de interesse primários são $\mu_i - \mu_j$ para $i = 1, \dots, j - 1$, a diferença entre cada nova média de tratamento μ_i e a média do controle μ_j , ou seja, queremos testar as hipóteses

$$\begin{cases} H_0: \mu_i = \mu_j \\ H_1: \mu_i \neq \mu_j \end{cases}$$

Teste de Comparações Múltiplas

Teste de Dunnett

O método de Dunnett é uma modificação do teste t usual. A menor diferença significativa neste caso é dada por

Tamanhos de amostras iguais
(dados balanceados)



$$d = d_{\alpha}(k, N - k) \sqrt{2 \times \frac{QME}{n}}$$

em que é um valor tabelado proposto por Dunnett, que depende do número de níveis (k) e dos graus de liberdade dos erros ($N-k$).

Tamanhos de amostras diferentes
(dados não balanceados)



$$d = d_{\alpha}(k, N - k) \sqrt{QME \left(\frac{1}{n_i} + \frac{1}{n_j} \right)}$$

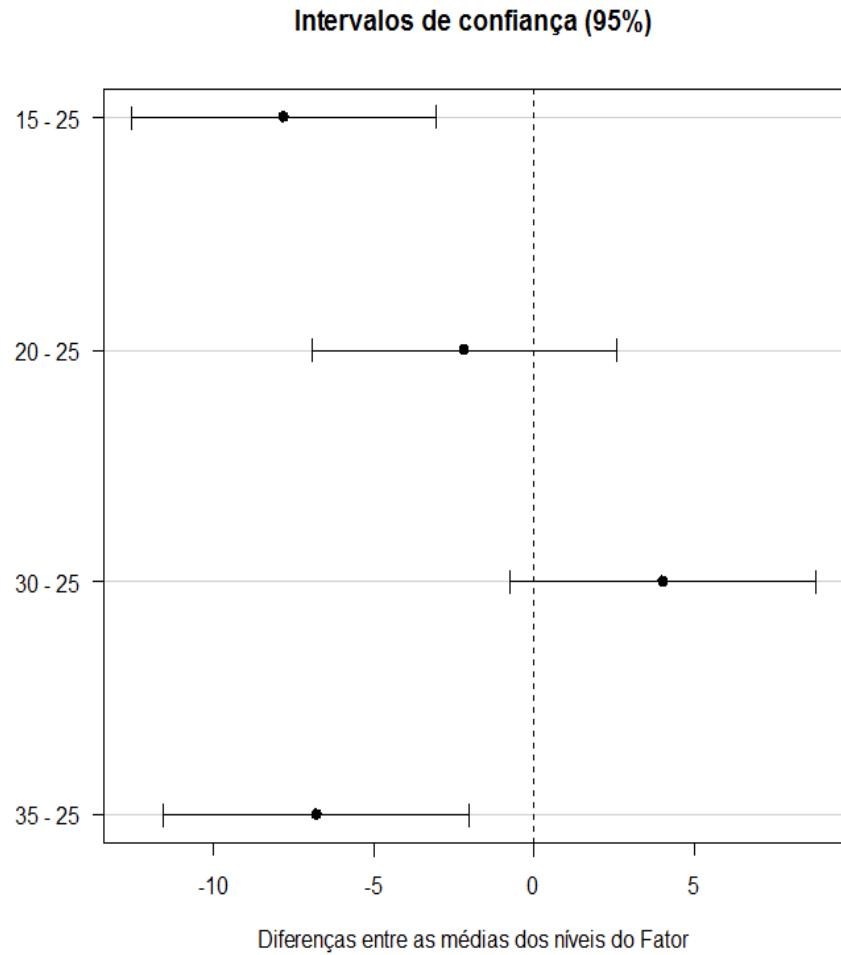
em que é um valor tabelado proposto por Dunnett, que depende do número de níveis (k) e dos graus de liberdade dos erros ($N-k$).

Exemplo

Considere o processo de produção de uma fibra sintética, onde o experimentador quer conhecer a influência da porcentagem de algodão na resistência da fibra. Para isto, foi realizado um experimento totalmente aleatorizado, onde diversos níveis de porcentagem de algodão foram avaliados com respeito à resistência da fibra. Um ponto importante na montagem do experimento é que para cada nível do fator (porcentagem de algodão), os outros fatores que influenciam o processo devem apresentar um padrão homogêneo de variabilidade. No experimento, tomamos 5 níveis para a porcentagem de algodão e 5 replicações.

Nível	Fator A					Soma	Média	Desvio Padrão
1	7	7	15	11	9	49	9,8	3,3
2	12	17	12	18	18	77	15,4	3,1
3	14	18	18	19	19	88	17,6	2,1
4	19	25	22	19	23	108	21,6	2,6
5	7	10	11	15	11	54	10,8	2,8
						$y_{..} = 376$	$\bar{y}_{..} = 15,04$	$s = 5,15$
								$s_{Entre} = 4,87$

Exemplo



$$d = d_{\alpha}(k, N - k) \sqrt{2 \times \frac{QME}{n}} = \\ = d_{0,05}(5,25 - 5) \sqrt{2 \times \frac{8,06}{5}} = 4,7582$$

Rejeitamos a igualdade entre dois níveis se:

$$|\bar{y}_{i\cdot} - \bar{y}_{j\cdot}| > 4,7582$$

Teste de Dunnett

DADOS DO PROCESSO

Alfa = 0,05
Hipótese Alternativa = Bilateral

Diferença entre Níveis	Média	LI	LS	P-Valor
15 - 25	-7,8	-12,56	-3,04	0,0010
20 - 25	-2,2	-6,96	2,56	0,56
30 - 25	4	-0,76	8,76	0,12
35 - 25	-6,8	-11,56	-2,04	0,0041

Teste de Comparações Múltiplas

Teste de HSU

Multiple Comparisons with the Best-
MCB

- ❑ Em algumas aplicações, a referência relevante (desconhecida) é o “melhor” tratamento, que é, o tratamento que tem maior valor de média (largest) ou menor (smallest), dependendo da análise de interesse.

- ❑ O teste proposto por Jason Hsu, tem como característica comparar todos os tratamentos com o melhor.

- ❑ Suponhamos que entre cinco tratamentos que estão sendo comparados, dois tratamentos são tão ruins que, a maioria dos pacientes que receberam um dos dois morreram dentro de um curto período de tempo.
- ❑ Possivelmente não é de interesse primordial saber qual desses dois tratamentos é pior, a inferência de que nenhum é melhor é suficiente.
- ❑ Suponhamos que o segundo melhor tratamento (entre os três restantes) é quase tão bom quanto o melhor tratamento verdadeiro.
- ❑ Assim, a inferência estatística que identifica ambos como praticamente o melhor pode ser de interesse, pois podem ter outras considerações que impactam na escolha do tratamento.
- ❑ Nestas situações todas as comparações duas a duas não são de interesse.

A principal questão aqui é
“Quais comparações são de interesse preliminar?”

Teste de Comparações Múltiplas

Teste de HSU

Podemos caracterizar as comparações de interesse principais nessas situações como “comparações múltiplas com o melhor”. Assim, se um efeito do tratamento maior é melhor, mesmo que o melhor tratamento seja desconhecido, podemos definir os parâmetros de interesse preliminar como

$$\max_{j=1,\dots,k} \mu_j - \mu_i, \quad i = 1, \dots, k \quad (*)$$

a diferença entre o efeito do melhor tratamento verdadeiro e cada um dos k efeitos do tratamento.

Maior é o melhor

Suponhamos que o maior efeito do tratamento implica em um tratamento melhor. Então os parâmetros

$$\mu_i - \max_{j \neq i} \mu_j, \quad i = 1, \dots, k$$

contém todas as informações que os parâmetros dados pela expressão (*)

Menor é o melhor

Naturalmente, se o menor efeito do tratamento implica no melhor tratamento, então por simetria os parâmetros de interesse preliminares são

$$\mu_i - \min_{j \neq i} \mu_j, \quad i = 1, \dots, k$$

Teste de Comparações Múltiplas

Teste de HSU

Maior é o melhor



O cálculo dos limites desses intervalos são obtidos usando as seguintes equações:

$$D_i^- = - \left[\bar{y}_i - \max_j (\bar{y}_j) - d_\alpha(k, N - k) \sqrt{2 \left(\frac{QME}{n} \right)} \right]^- \quad \text{Limite Inferior}$$
$$D_i^+ = \left[\bar{y}_i - \max_j (\bar{y}_j) + d_\alpha(k, N - k) \sqrt{2 \left(\frac{QME}{n} \right)} \right]^+ \quad \text{Limite Superior}$$

Se o intervalo ($D_i^- ; D_i^+$) assumir somente valores positivos, consideramos que o i-ésimo nível do fator é o melhor.

Menor é o melhor



O cálculo dos limites desses intervalos são obtidos usando as seguintes equações:

$$D_i^- = - \left[\bar{y}_i - \min_{j \neq i} (\bar{y}_j) - d_\alpha(k, N - k) \sqrt{2 \left(\frac{QME}{n_i} \right)} \right]^- \quad \text{Limite Inferior}$$
$$D_i^+ = \left[\bar{y}_i - \min_{j \neq i} (\bar{y}_j) + d_\alpha(k, N - k) \sqrt{2 \left(\frac{QME}{n_i} \right)} \right]^+ \quad \text{Limite Superior}$$

Se o intervalo ($D_i^- ; D_i^+$) assumir somente valores negativos, consideramos que o i-ésimo nível do fator é o melhor.

Para simplificar a análise e disposição dos resultados em um gráfico, realizamos a seguinte transformação dos limites dos intervalos de confiança. Para cada valor de D_i , calculamos:

$$[D_i^-]' = \min\{0, D_i^-\} = \begin{cases} D_i^-, & \text{se } x \text{ é negativo} \\ 0, & \text{caso contrário} \end{cases}$$
$$[D_i^+]' = \max\{0, D_i^+\} = \begin{cases} D_i^+, & \text{se } x \text{ é positivo} \\ 0, & \text{caso contrário} \end{cases}$$

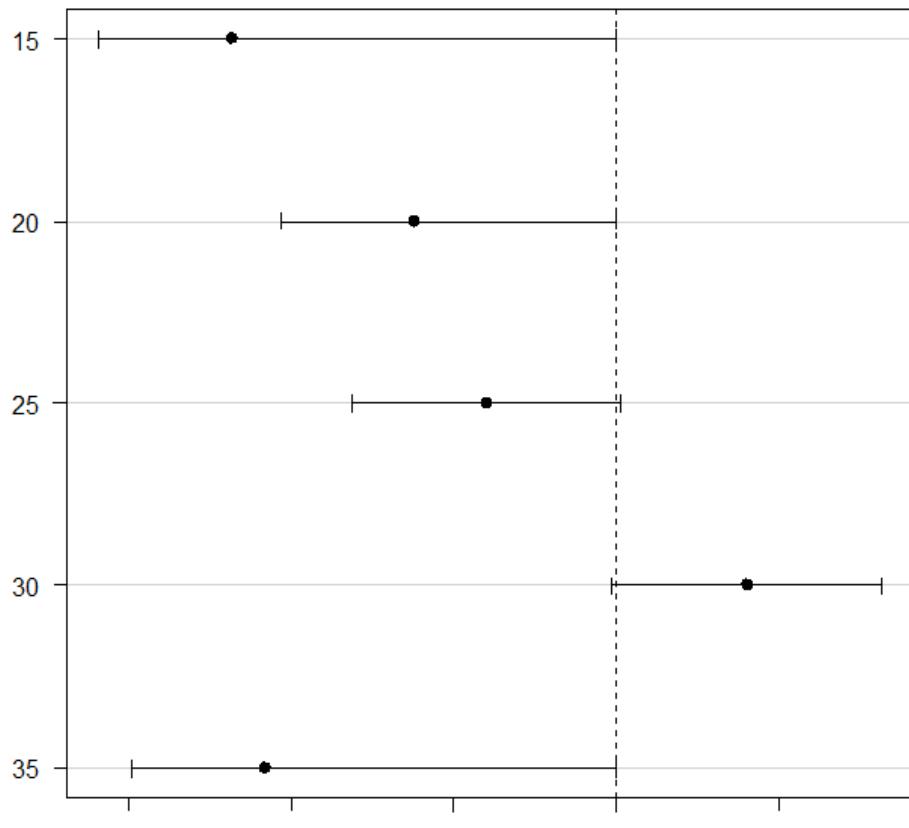
Exemplo

Considere o processo de produção de uma fibra sintética, onde o experimentador quer conhecer a influência da porcentagem de algodão na resistência da fibra. Para isto, foi realizado um experimento totalmente aleatorizado, onde diversos níveis de porcentagem de algodão foram avaliados com respeito à resistência da fibra. Um ponto importante na montagem do experimento é que para cada nível do fator (porcentagem de algodão), os outros fatores que influenciam o processo devem apresentar um padrão homogêneo de variabilidade. No experimento, tomamos 5 níveis para a porcentagem de algodão e 5 replicações.

Nível	Fator A					Soma	Média	Desvio Padrão
1	7	7	15	11	9	49	9,8	3,3
2	12	17	12	18	18	77	15,4	3,1
3	14	18	18	19	19	88	17,6	2,1
4	19	25	22	19	23	108	21,6	2,6
5	7	10	11	15	11	54	10,8	2,8
						$y_{..} = 376$	$\bar{y}_{..} = 15,04$	$s = 5,15$
								$s_{Entre} = 4,87$

Exemplo

Intervalos de confiança (95%)



Diferenças entre as médias dos níveis do Fator

$$d = d_{\alpha}(k, N - k) \sqrt{2 \times \frac{QME}{n}} = \\ = 2,305 \sqrt{2 \times \frac{8,06}{5}} = 4,138$$

Para o nível 15, temos:

$$D_i^- = -[-11,8 - 4,138]^- = -[-15,938]^- = -15,938$$

$$D_i^+ = [-11,8 + 4,138]^+ = [-7,662]^+ = 0$$

Repetindo para os demais níveis, obtemos:

TESTE HSU

DADOS DO PROCESSO

TESTE: Maior é o melhor

Nível	Média	LI	LS
15	-11,8	-15,94	0
20	-6,2	-10,34	0
25	-4	-8,14	0,14
30	4	-0,14	8,14
35	-10,8	-14,94	0

Análise de Variância

ANOVA

Efeito fixo

2 Fator Fatores

ANOVA Dois Fatores

ANOVA

Ferramenta para análise do comportamento de diversos tratamentos para dois fatores, aplicados ao processo e/ou produto

O comportamento é avaliado com relação a uma resposta do processo e/ou produto

Sejam A o primeiro fator, com a níveis e B o segundo fator, com b níveis e r as réplicas do experimento.

Fator A	Fator B				Média
	1	2	...	b	
1	y_{111}, \dots, y_{11r}	y_{121}, \dots, y_{12r}	...	y_{1b1}, \dots, y_{1br}	$\bar{y}_{1..}$
2	y_{211}, \dots, y_{21r}	y_{221}, \dots, y_{22r}	...	y_{2b1}, \dots, y_{2br}	$\bar{y}_{2..}$
:	:	:	:	:	:
a	y_{a11}, \dots, y_{a1r}	y_{a21}, \dots, y_{a2r}	...	y_{ab1}, \dots, y_{abr}	$\bar{y}_{a..}$
Média	$\bar{y}_{.1.}$	$\bar{y}_{.2.}$...	$\bar{y}_{.b.}$	$\bar{y}_{...}$

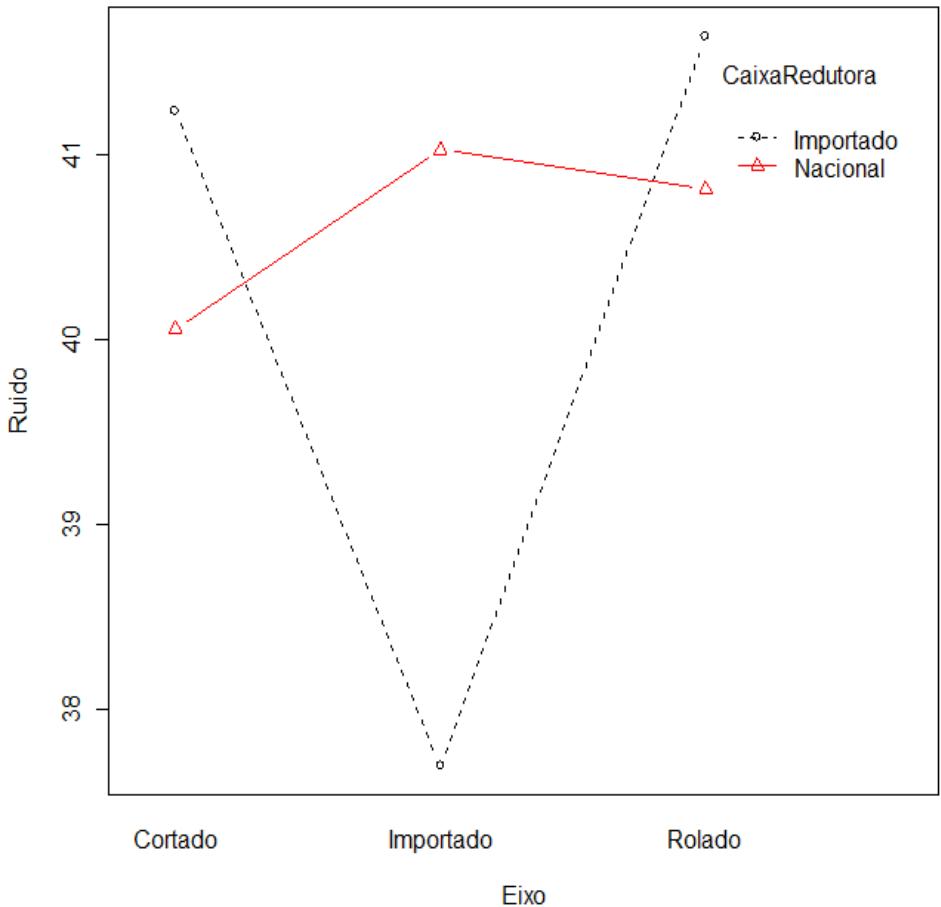
Exemplo

Uma empresa que produz limpadores de para-brisas para automóveis quer saber como os fatores Tipo de Caixa Redutora e Tipo de Eixo, utilizados na fabricação dos motores que acionam os limpadores, influenciam o ruído produzido, quando da utilização destes. Para isso realizamos um experimento com 54 motores, com 3 tipos de Eixo (Rolado, Cortado e Importado) e 2 tipos de Caixas Redutora (Nacional e Importada). Para cada motor (unidade experimental) medimos o ruído.

Tipos de Caixa Redutora	Tipos de Eixo								
	Rolado			Cortado			Importado		
Nacional	42,1	42	40,3	38,2	37,4	37	40,9	40,7	39,4
	38,9	38,9	43,7	42,3	41,3	42,1	42	41,4	41,3
	41	40,1	40,3	40,5	41,3	40,4	40,6	41,3	41,6
Importada	39,6	40,2	48,4	41,3	46,8	40,3	39,6	36,9	39,9
	40,9	41	41	40,5	39,9	39,3	38,1	38	36,2
	39,9	41	42,7	41,3	40,1	41,6	36,7	37,2	36,7

Exemplo

Gráfico de Interações



Conclusão:

- A interação entre os fatores está associada à mudança de comportamento de um fator nos diferentes níveis do outro fator, com relação à característica de interesse.
- Na figura observamos que quando a caixa redutora é nacional (linha vermelha), os três níveis de eixo (Cortado, Importado e Rolado) não provocam mudança significativa no ruído do motor.
- Porém, quando a caixa redutora é importada (linha pontilhada), existe diferença de ruído entre os três tipos de eixo.
- Neste caso, para um eixo importado temos menor ruído.
- Desta forma, evidenciamos uma interação entre os fatores (caixa redutora e eixo) na característica de interesse (ruído).
- Dependendo do tipo de caixa redutora, o comportamento do eixo, com respeito ao ruído, é diferente. Essa diferença caracteriza o que denominamos interação.

Modelo da ANOVA

O modelo com dois fatores balanceados e com efeitos cruzados com interação é dado por

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \tau_{ij} + \varepsilon_{ijk},$$

com,

$i = 1, 2, \dots, a$ Fator A

$j = 1, 2, \dots, b$ Fator B

$k = 1, \dots, r$ Réplica

Em que

- μ : média geral dos dados
- α_i : efeito do nível i do fator A
- β_j : efeito do nível j do fator B
- τ_{ij} : efeito da interação ij entre os fatores
- ε_{ijk} : componente aleatória do erro

Notações:

- $y_{i..} = \sum_{j=1}^b \sum_{k=1}^r y_{ijk}$ = soma das observações do nível i do fator A
- $\bar{y}_{i..} = \frac{y_{i..}}{br}$ = média das observações do nível i do fator A
- $y_{..} = \sum_{i=1}^n \sum_{j=1}^{n_i} y_{ij}$ = soma de todas as observações
- $\bar{y}_{...} = \frac{1}{b a r} \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^r y_{ijk}$ = média geral das observações
- Analogamente, $\bar{y}_{.j.} = \frac{1}{a r} \sum_{i=1}^a \sum_{k=1}^r y_{ijk}$, $\bar{y}_{ij.} = \frac{1}{r} \sum_{k=1}^r y_{ijk}$

Análise de variância - ANOVA

Hipóteses



$\varepsilon_{ijk} \sim N(0, \sigma^2)$ independentes.



$y_{ijk} \sim N(\mu + \alpha_i + \beta_j + \tau_{ij}, \sigma^2)$ independentes para $i = 1, \dots, k$.

Teste de Hipóteses



Efeito do Fator A

$$\begin{cases} H_0: \alpha_1 = \alpha_2 = \dots = \alpha_a \\ H_1: \text{pelo menos um é diferente.} \end{cases}$$



Efeito do Fator B

$$\begin{cases} H_0: \beta_1 = \beta_2 = \dots = \beta_b \\ H_1: \text{pelo menos um é diferente.} \end{cases}$$



Efeito da Interação $A \times B$

$$\begin{cases} H_0: \tau_{ij} = 0 \text{ para todos os valores de } i \text{ e } j \\ H_1: \tau_{ij} \neq 0 \end{cases}$$

Decomposição da Soma de Quadrados

A soma de quadrados total é definida como medida da variabilidade total dos dados

$$\sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^r (y_{ijk} - \bar{y}_{...})^2 = b r \sum_{i=1}^a (\bar{y}_{i..} - \bar{y}_{...})^2 + a r \sum_{j=1}^b (\bar{y}_{.j.} - \bar{y}_{...})^2 + r \sum_{i=1}^a \sum_{j=1}^b (\bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}_{...})^2 + \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^r (y_{ijk} - \bar{y}_{ij.})^2$$

$$SQT = SQA + SQB + SQAB + QE$$

Modelo da ANOVA: Graus de Liberdade e Estimativas da Variância

QMA



$$E(QM_A) = \sigma^2 + \frac{b r}{a - 1} \sum_{i=1}^a (\alpha_i)^2$$

QMB



$$E(QM_B) = \sigma^2 + \frac{a r}{b - 1} \sum_{j=1}^b (\beta_j)^2$$

QMAB



$$E(QM_{AB}) = \sigma^2 + \frac{r}{(b - 1)(a - 1)} \sum_{i=1}^a \sum_{j=1}^b (\tau_{ij})^2$$

QME



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

Efeito	Graus de Liberdade
Fator A	$a - 1$
Fator B	$b - 1$
Interação $A \times B$	$(a - 1)(b - 1)$
Erro	$ab(r - 1)$
Total	$abr - 1$

Análise de variância - ANOVA

Através do teorema de Cochran, garantimos, sob H_0 , a independência das somas de quadrados

Efeito do Fator A

$$\begin{cases} H_0: \alpha_1 = \alpha_2 = \cdots = \alpha_a \\ H_1: \text{pelo menos um é diferente.} \end{cases}$$

→ $\frac{SQ_A}{\sigma^2} \sim \chi_{(a-1)}^2 \quad \text{e} \quad \frac{SQ_E}{\sigma^2} \sim \chi_{(a b (r-1))}^2,$

$$F_0 = \frac{\frac{SQ_A}{(\sigma^2) (a-1)}}{\frac{SQ_E}{\sigma^2 a b (r-1)}} = \frac{QM_A}{QM_E} \sim F(a-1; a b (r-1)),$$

Efeito do Fator B

$$\begin{cases} H_0: \beta_1 = \beta_2 = \cdots = \beta_b \\ H_1: \text{pelo menos um é diferente.} \end{cases}$$

→ $\frac{SQ_B}{\sigma^2} \sim \chi_{(b-1)}^2 \quad \text{e} \quad \frac{SQ_E}{\sigma^2} \sim \chi_{(a b (r-1))}^2,$

$$F_0 = \frac{\frac{SQ_B}{(\sigma^2) (b-1)}}{\frac{SQ_E}{\sigma^2 (a b (r-1))}} = \frac{QM_B}{QM_E} \sim F(b-1; a b (r-1)),$$

Efeito da Interação $A \times B$

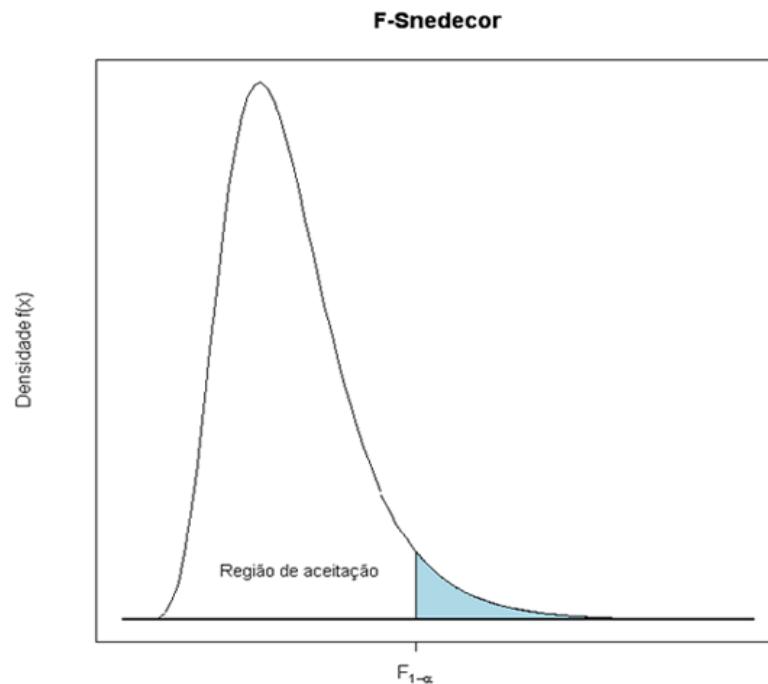
$$\begin{cases} H_0: \tau_{ij} = 0 \text{ para todos os valores de } i \text{ e } j \\ H_1: \tau_{ij} \neq 0 \end{cases}$$

→ $\frac{SQ_{AB}}{\sigma^2} \sim \chi_{(a-1)(b-1)}^2 \quad \text{e também,} \quad \frac{SQ_E}{\sigma^2} \sim \chi_{(a b (r-1))}^2,$

$$F_0 = \frac{\frac{SQ_{AB}}{(\sigma^2) (a-1)(b-1)}}{\frac{SQ_E}{(\sigma^2) (a b (r-1))}} = \frac{QM_{AB}}{QM_E} \sim F((a-1)(b-1); (ab(r-1)))$$

Teste da ANOVA - Dois Fatores

Fator	Soma de Quadrados	Graus de Liberdade	Quadrados Médios	F_0
Fator A	SQA	$a - 1$	$QMA = \frac{SQA}{a - 1}$	$F_A = \frac{QMA}{QME}$
Fator B	SQB	$b - 1$	$QMB = \frac{SQB}{b - 1}$	$F_B = \frac{QMB}{QME}$
Interação ($A \times B$)	SQAB	$(a - 1)(b - 1)$	$QMA = \frac{SQAB}{(a - 1)(b - 1)}$	$F_{AB} = \frac{QMAB}{QME}$
Erro	SQE	$ab(r - 1)$	$QME = \frac{SQE}{ab(r - 1)}$	
Total	SQT	$abr - 1$		



Exemplo

Uma empresa que produz limpadores de para-brisas para automóveis quer saber como os fatores Tipo de Caixa Redutora e Tipo de Eixo, utilizados na fabricação dos motores que acionam os limpadores, influenciam o ruído produzido, quando da utilização destes. Para isso realizamos um experimento com 54 motores, com 3 tipos de Eixo (Rolado, Cortado e Importado) e 2 tipos de Caixas Redutora (Nacional e Importada). Para cada motor (unidade experimental) medimos o ruído.

Tipos de Caixa Redutora	Tipos de Eixo							
	Rolado			Cortado			Importado	
Nacional	42,1	42	40,3	38,2	37,4	37	40,9	40,7
	38,9	38,9	43,7	42,3	41,3	42,1	42	41,4
	41	40,1	40,3	40,5	41,3	40,4	40,6	41,3
Importada	39,6	40,2	48,4	41,3	46,8	40,3	39,6	36,9
	40,9	41	41	40,5	39,9	39,3	38,1	38
	39,9	41	42,7	41,3	40,1	41,6	36,7	36,7

ANOVA

DADOS DO PROCESSO

Tabela da Anova					
	G.L.	Soma de Quadrados	Quadrado Médio	Estat. F	P-valor
Eixo	2	32,67	16,33	4,67	0,014
CaixaRedutora	1	2,62	2,62	0,75	0,39
Eixo:CaixaRedutora	2	56,33	28,16	8,056	0,001
Resíduos	48	167,81	3,50		

Estimação dos Parâmetros do Modelo

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \tau_{ij} + \varepsilon_{ijk},$$

com, $i = 1, 2, \dots, a$
 $j = 1, 2, \dots, b$

Se assumirmos que os erros ε_{ijk} são normalmente distribuídos, obtemos

$$\bar{y}_{i..} \sim N\left(\mu + \alpha_i, \frac{\sigma^2}{br}\right) \text{ e}$$

$$\bar{y}_{.j.} \sim N\left(\mu + \beta_j, \frac{\sigma^2}{ar}\right)$$

- $\hat{\mu} = \bar{y}_{...}$ estimador da média geral
- $\hat{\alpha}_i = \bar{y}_{i..} - \bar{y}_{...}$ e $\hat{\beta}_j = \bar{y}_{.j.} - \bar{y}_{...}$ estimador dos efeitos
- $\mu_i = \mu + \alpha_i$ e $\mu_{.j.} = \mu + \beta_j$ as médias do i-ésimo e j-ésimo níveis respectivamente.

Intervalos com confiança de $(1 - \alpha)100\%$ para μ_i e μ_j são

$$\bar{y}_{i..} - t_{\left(1 - \frac{\alpha}{2}, ab(r-1)\right)} \times \sqrt{\frac{QME}{br}} \leq \mu_i \leq \bar{y}_{i..} + t_{\left(1 - \frac{\alpha}{2}, ab(r-1)\right)} \times \sqrt{\frac{QME}{br}}$$

$$\bar{y}_{.j.} - t_{\left(1 - \frac{\alpha}{2}, ab(r-1)\right)} \times \sqrt{\frac{QME}{ar}} \leq \mu_{.j.} \leq \bar{y}_{.j.} + t_{\left(1 - \frac{\alpha}{2}, ab(r-1)\right)} \times \sqrt{\frac{QME}{ar}}$$

$$\frac{\frac{(\bar{y}_{i..} - \mu_i)}{\sqrt{\sigma^2 / br}}}{\sqrt{\frac{ab(r-1)QME}{\sigma^2 / ar}}} = \frac{(\bar{y}_{i..} - \mu_i)}{\sqrt{QME / br}}$$

tem distribuição t-Student com $ab(r-1)$ graus de liberdade.

Estimação dos Parâmetros do Modelo

Vamos construir o Intervalo com confiança de $(1 - \alpha)100\%$ para a diferença entre as médias de dois níveis, para isto,

$$\frac{(\bar{y}_{i..} - \bar{y}_{l..}) - (\alpha_i - \alpha_l)}{\sqrt{2\sigma^2/br}} \sim N(0,1)$$

$$\frac{\frac{\bar{y}_{i..} - \bar{y}_{l..} - (\alpha_i - \alpha_l)}{\sqrt{2\sigma^2/br}}}{\sqrt{\frac{QME ab(r-1)}{ab(r-1)}} / \sqrt{\frac{\sigma^2}{ab(r-1)}}} = \frac{\bar{y}_{i..} - \bar{y}_{l..} - (\alpha_i - \alpha_l)}{\sqrt{\frac{2}{br} QME}} \sim t_{(ab(r-1))}$$

Portanto, o Intervalo com confiança de $(1 - \alpha)100\%$ para a diferença entre as médias de dois níveis do Fator A é

$$(\bar{y}_{i..} - \bar{y}_{l..}) - \Delta \leq \mu_{i..} - \mu_{l..} \leq (\bar{y}_{i..} - \bar{y}_{l..}) + \Delta$$

em que $\Delta = t_{(1-\frac{\alpha}{2}; ab(r-1))} \times \sqrt{QME \left(\frac{2}{br} \right)}$

Similarmente, o Intervalo com confiança de $(1 - \alpha)100\%$ para a diferença entre as médias de dois níveis do Fator B é

$$(\bar{y}_{j..} - \bar{y}_{l..}) - \Delta \leq \mu_{j..} - \mu_{l..} \leq (\bar{y}_{j..} - \bar{y}_{l..}) + \Delta$$

em que $\Delta = t_{(1-\frac{\alpha}{2}; ab(r-1))} \times \sqrt{QME \left(\frac{2}{ar} \right)}$

Exemplo

Uma empresa que produz limpadores de para-brisas para automóveis quer saber como os fatores Tipo de Caixa Redutora e Tipo de Eixo, utilizados na fabricação dos motores que acionam os limpadores, influenciam o ruído produzido, quando da utilização destes. Para isso realizamos um experimento com 54 motores, com 3 tipos de Eixo (Rolado, Cortado e Importado) e 2 tipos de Caixas Redutora (Nacional e Importada). Para cada motor (unidade experimental) medimos o ruído.

Tipos de Caixa Redutora	Tipos de Eixo								
	Rolado			Cortado			Importado		
Nacional	42,1	42	40,3	38,2	37,4	37	40,9	40,7	39,4
	38,9	38,9	43,7	42,3	41,3	42,1	42	41,4	41,3
	41	40,1	40,3	40,5	41,3	40,4	40,6	41,3	41,6
Importada	39,6	40,2	48,4	41,3	46,8	40,3	39,6	36,9	39,9
	40,9	41	41	40,5	39,9	39,3	38,1	38	36,2
	39,9	41	42,7	41,3	40,1	41,6	36,7	37,2	36,7

INTERVALO DE CONFIANÇA DAS MÉDIAS

DADOS DO PROCESSO

Exemplo

Eixo	Média	Desvio Padrão	Limite Inferior	Limite Superior
Cortado	40,64	2,12	39,64	41,65
Importado	39,36	2,12	38,36	40,36
Rolado	41,22	2,12	40,22	42,22

Caixa Redutora	Média	Desvio Padrão	Limite Inferior	Limite Superior
Importado	40,19	2,12	39,37	41,01
Nacional	40,63	2,12	39,81	41,45

Gráfico de Intervalos

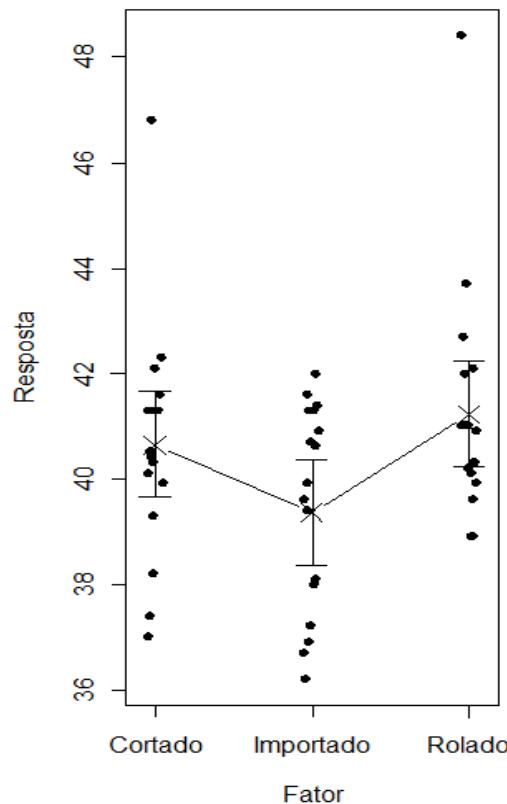
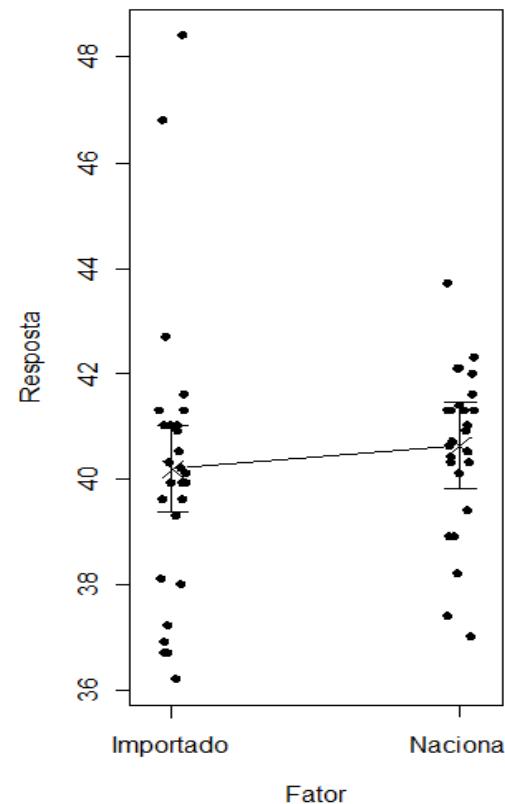


Gráfico de Intervalos



Análise de Resíduos

Hipóteses

- $\varepsilon_{ijk} \sim N(0, \sigma^2)$ independentes e identicamente distribuídos;
- Modelo para os dados:

$$y_{ijk} = \mu + \alpha_i + \beta_j + \tau_{ij} + \varepsilon_{ijk}$$

Na prática, precisamos verificar se estas suposições são válidas. Violações nestas suposições são verificadas através dos resíduos

O resíduo para a k-ésima observação do nível i do Fator A e nível j do Fator B

$$\hat{e}_{ijk} = y_{ijk} - \hat{y}_{ijk}$$

Em que \hat{y}_{ijk} é uma estimativa da observação y_{ijk}

$$\hat{y}_{ijk} = \hat{\mu} + \hat{\alpha}_i + \hat{\beta}_j + \hat{\tau}_{ij} = \bar{y}_{ij}.$$

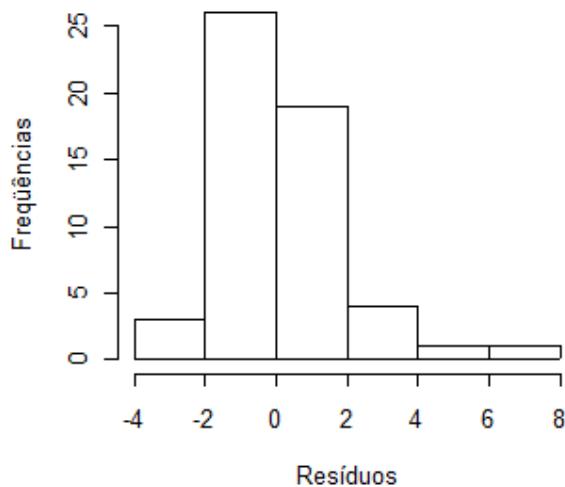
Exemplo

Uma empresa que produz limpadores de para-brisas para automóveis quer saber como os fatores Tipo de Caixa Redutora e Tipo de Eixo, utilizados na fabricação dos motores que acionam os limpadores, influenciam o ruído produzido, quando da utilização destes. Para isso realizamos um experimento com 54 motores, com 3 tipos de Eixo (Rolado, Cortado e Importado) e 2 tipos de Caixas Redutora (Nacional e Importada). Para cada motor (unidade experimental) medimos o ruído.

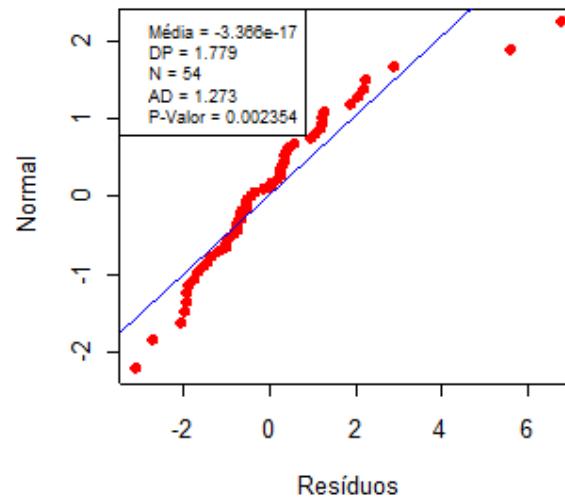
Tipos de Caixa Redutora	Tipos de Eixo								
	Rolado			Cortado			Importado		
Nacional	42,1	42	40,3	38,2	37,4	37	40,9	40,7	39,4
	38,9	38,9	43,7	42,3	41,3	42,1	42	41,4	41,3
	41	40,1	40,3	40,5	41,3	40,4	40,6	41,3	41,6
Importada	39,6	40,2	48,4	41,3	46,8	40,3	39,6	36,9	39,9
	40,9	41	41	40,5	39,9	39,3	38,1	38	36,2
	39,9	41	42,7	41,3	40,1	41,6	36,7	37,2	36,7

Exemplo

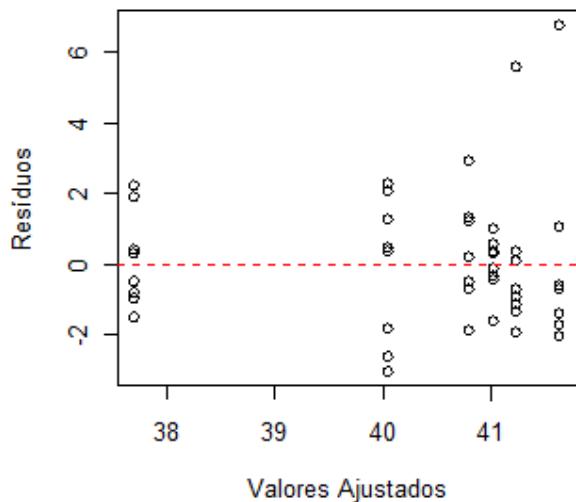
Histograma dos Resíduos



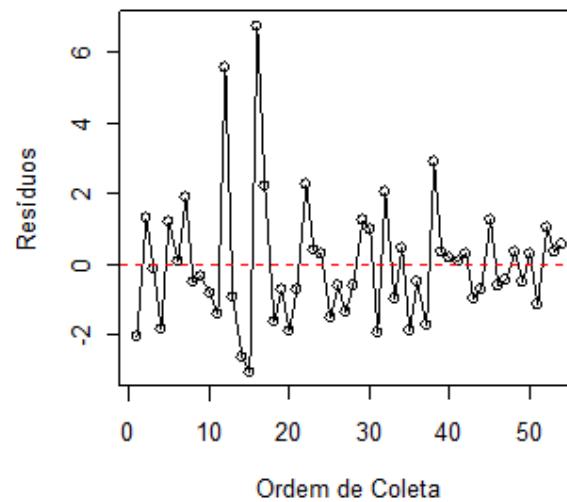
Papel de Probabilidade



Resíduos x Valores Ajustados



Resíduos x Ordem de Coleta



Influência do R^2 na ANOVA

Uma maneira de verificarmos se o modelo ajustado é adequado é olharmos o resultado do coeficiente de determinação (R^2). Este coeficiente mede o quanto a variável resposta é explicada pelo modelo.

A expressão para calcular o R^2

$$R^2 = 1 - \frac{SQE}{SQT}$$

Em uma análise de variância com efeito fixo, estamos interessado em determinar se existe diferença entre os níveis dos fatores. Aqui, não temos interesse em utilizar o modelo para previsão. Assim, a adequabilidade do modelo linear não é crucial para aplicação da ANOVA. Para comprovarmos a afirmação, realizamos um estudo de simulação, conforme abaixo:

- Utilizamos um fator com 4 níveis, com 10 réplicas em cada nível;
- Para cada nível geramos uma distribuição Normal com médias 12; 12; 3; 11,5 e desvio padrão 1;
- Logo após, realizamos uma ANOVA e registramos os valores de p (*p*-valor) e do R^2 ;
- Repetimos o procedimento acima 10.000 vezes, obtendo os seguinte resultados:
 - ✓ Média do R^2 = 0,4339
 - ✓ Porcentagem dos p-valores maiores que 0,05 é zero

Conclusão: Portanto, a não adequabilidade do modelo ($R^2 < 0,70$) não influencia de forma significativa o resultado do teste *F* da ANOVA.

Análise de Regressão

Regressão Linear Simples

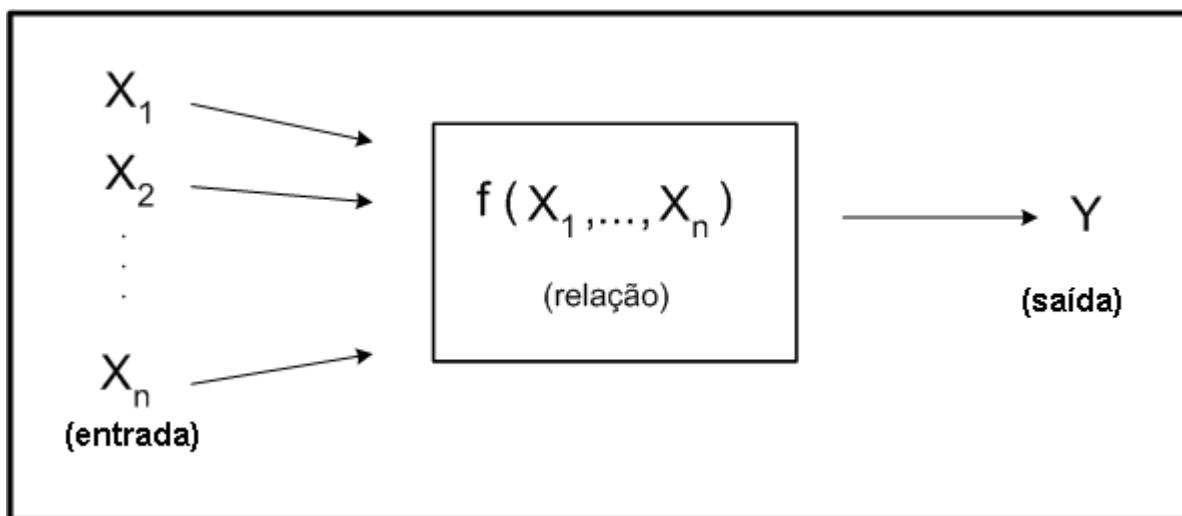
Introdução a regressão

Análise de Regressão

Possibilita encontrar uma relação razoável entre as variáveis de entrada e saída, por meio de relações empíricas.

A utilização desta abordagem necessita de coleta de dados e do uso de métodos estatísticos de Análise de Regressão Linear

A coleta de dados permite conhecer a natureza da relação entre as variáveis e realizar estudos capazes de acomodar situações inesperadas, como por exemplo, variabilidade na matéria prima, temperatura ambiente, máquina e operadores



Modelos de Regressão

Objetivos

Predição: Uma vez que esperamos que grande parte da variação da variável de saída seja explicada pelas variáveis de entrada, podemos utilizar o modelo para obter valores de Y correspondentes a valores de X que não estavam entre os dados. Esse procedimento é chamado de predição e, em geral, usamos valores de X que estão dentro do intervalo de variação estudado.

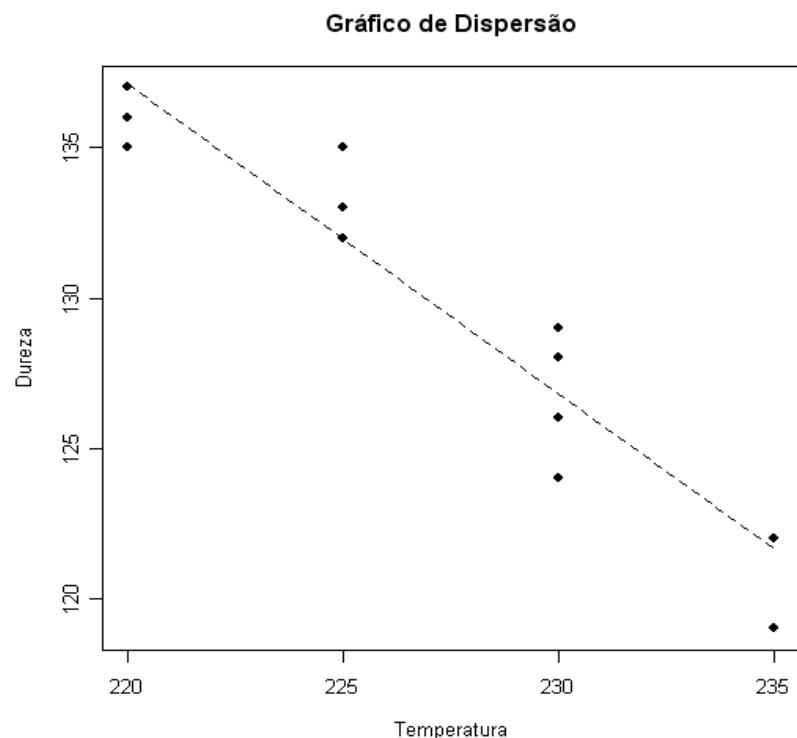
Seleção de variáveis: Frequentemente, não se tem ideia de quais são as variáveis que afetam significativamente a variação de Y. Para responder a esse tipo de questão, estudos são realizados com um grande número de variáveis.

Estimação de parâmetros: Dado um modelo e um conjunto de dados referente às variáveis resposta e preditoras, estimar parâmetros ou ajustar um modelo aos dados significa obter valores ou estimativas para os parâmetros, por algum processo, tendo por base o modelo e os dados observados.

Inferência: O ajuste de um modelo de regressão em geral tem por objetivos básicos, além de estimar os parâmetros, realizar inferências sobre eles, tais como, testes de hipóteses e intervalos de confiança.

Motivação 1

Em problemas de tratamento térmico deseja-se estabelecer uma relação entre a temperatura da estufa e uma característica da qualidade (Dureza). Determinar os valores de temperatura em C que “otimizam” a performance do processo de tratamento térmico em relação a estrutura metalográfica do material, avaliada em relação de dureza em HB.



Observação	Dureza (HB)	Temperatura (°C)	Observação	Dureza (HB)	Temperatura (°C)
1	137	220	11	128	230
2	137	220	12	124	230
3	137	220	13	126	230
4	136	220	14	129	230
5	135	220	15	126	230
6	135	225	16	122	235
7	133	225	17	122	235
8	132	225	18	122	235
9	133	225	19	119	235
10	133	225	20	122	235

Modelo Estatístico

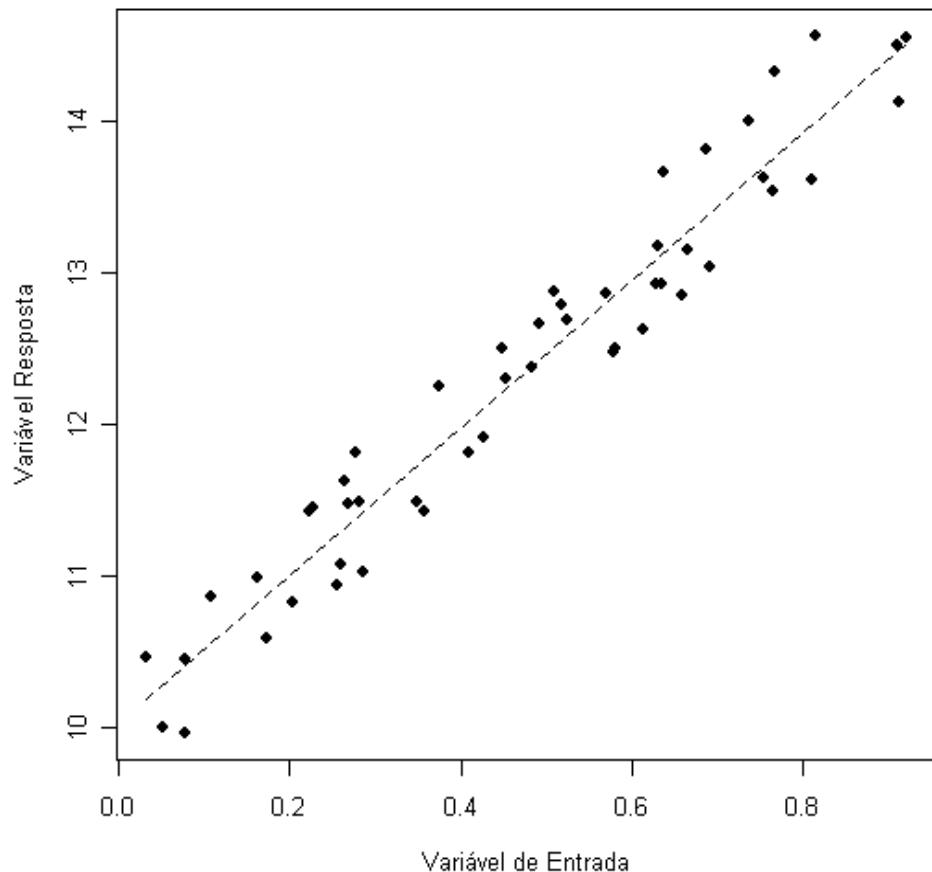
Propomos a seguinte relação

$$E(Y|x) = \beta_0 + \beta_1 x$$

Em que:

- β_0 indica o ponto que a reta intercepta o eixo y, isto é, $x = 0$;
- β_1 representa a inclinação da reta

Gráfico de Dispersão



Modelo Estatístico

Definição

Dados n pares de valores, $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$, se y é função linear de x , pode-se estabelecer uma regressão linear simples, cujo modelo estatístico

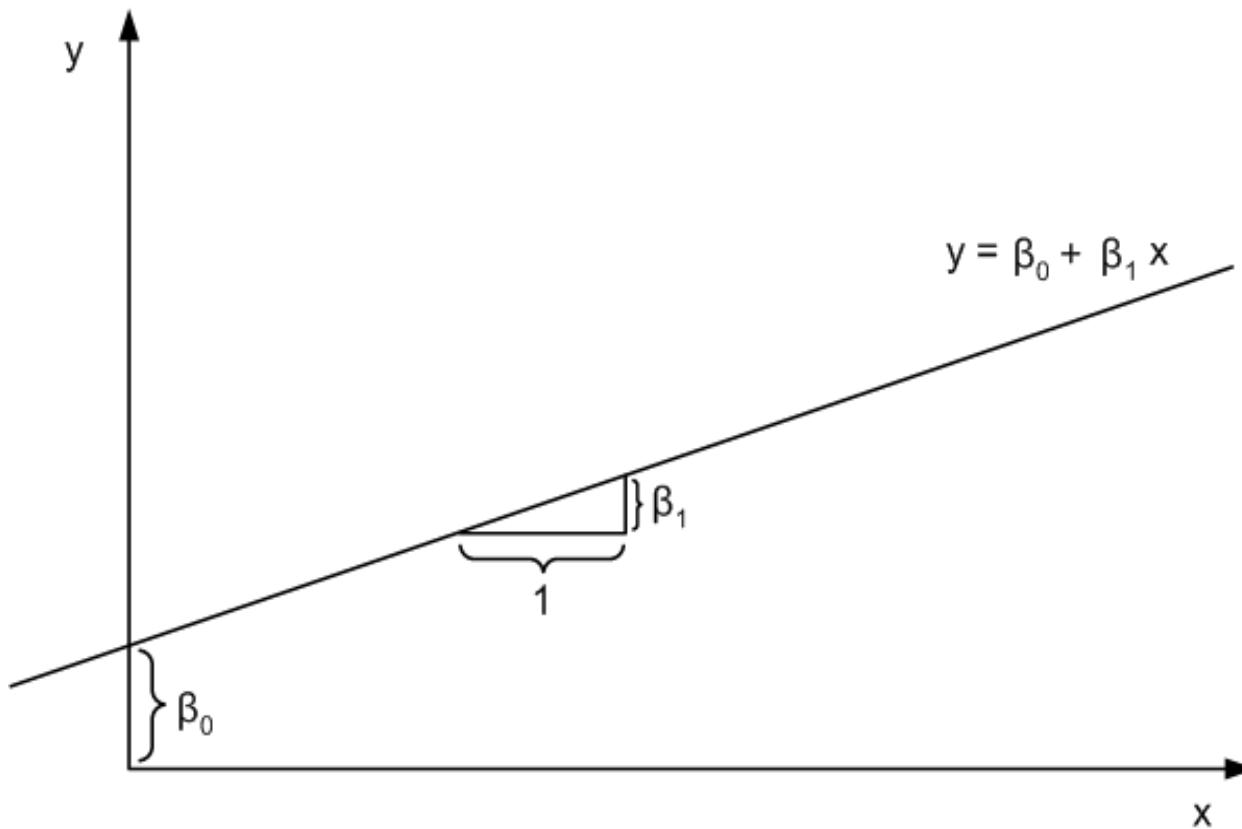
$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

para $i = 1, \dots, n.$

Neste modelo:

- Y_i é uma variável aleatória e representa o valor da variável resposta;
- x_i são as variáveis explicativas;
- ε_i uma variável aleatória que representa o erro experimental;
- β_0 e β_1 são os parâmetros do modelo, que serão estimados, e que definem a reta de regressão;
- n é o tamanho da amostra.

Representação da reta de regressão



Um ponto negativo na **Definição** é que o modelo de regressão linear simples não acomoda impactos de erros experimentais (variação de matéria prima), de erros de medida, entre outras inúmeras fontes de variabilidade, tornando-se inadequado nestes casos.

Suposições do Modelo

Modelo

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad \text{para } i = 1, \dots, n.$$

A relação matemática entre y e x é linear

Os valores de x são fixos (ou controlados)

A média do erro é nula, ou seja, $E(\varepsilon_i) = 0$

Para um dado valor de x , a variância de ε_i é sempre σ^2 , ou seja,
podemos dizer que o erro é homocedástico (tem variância constante)
 $Var(\varepsilon_i) = Var(Y_i) = \sigma^2$

O erro de uma observação é independente do erro de outra
observação (erros não correlacionados), isto é,

$$Cov(\varepsilon_i, \varepsilon_j) = 0, \text{ para } i \neq j$$

Frequentemente, supomos que os erros tem distribuição Normal

Estimação dos parâmetros

Método dos Mínimos Quadrados

Estimar os parâmetros β_0 e β_1 de modo que os desvios
 $\varepsilon_i = Y_i - [\beta_0 + \beta_1 x_i]$
sejam mínimos

Consiste em minimizar L , que é a soma dos quadrados dos desvios

$$L = \sum_{i=1}^m \varepsilon_i^2 = \sum_{i=1}^n [Y_i - \beta_0 - \beta_1 x_i]^2$$

Lembrando que $\sum_{i=1}^n \varepsilon_i = 0$

Passos para minimizar L

Derivar a função L , em relação aos parâmetros

Substituir β_0 e β_1 por $\hat{\beta}_0$ e $\hat{\beta}_1$

Igualar as derivadas à zero

Resolver o sistema de equações, em relação as variáveis $\hat{\beta}_0$ e $\hat{\beta}_1$

Estimação dos parâmetros

Estimativas
de Mínimos
Quadrados

$$\hat{\beta}_0 = \frac{1}{n} \sum_i^n y_i - \hat{\beta}_1 \frac{1}{n} \sum_i^n x_i = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n x_i y_i - n \bar{x} \bar{y}}{\sum_{i=1}^n x_i^2 - \frac{(\sum_{i=1}^n x_i)^2}{n}} = \frac{S_{xy}}{S_{xx}}$$

A reta de regressão ajustada é dada por

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i , \quad i = 1, 2, \dots, n.$$

Exemplo

Em problemas de tratamento térmico deseja-se estabelecer uma relação entre a temperatura da estufa e uma característica da qualidade (Dureza). Determinar os valores de temperatura em C que “otimizam” a performance do processo de tratamento térmico em relação a estrutura metalográfica do material, avaliada em relação de dureza em HB.

Observação	Temperatura (x)	Dureza (y)	x^2	y^2	$x \times y$
1	220	137	48.400	18.769	30.140
2	220	137	48.400	18.769	30.140
3	220	137	48.400	18.769	30.140
4	220	136	48.400	18.496	29.920
5	220	135	48.400	18.225	29.700
6	225	135	50.625	18.225	30.375
7	225	133	50.625	17.689	29.925
8	225	132	50.625	17.424	29.700
9	225	133	50.625	17.689	29.925
10	225	133	50.625	17.689	29.925
11	230	128	52.900	16.384	29.440
12	230	124	52.900	15.376	28.520
13	230	126	52.900	15.876	28.980
14	230	129	52.900	16.641	29.670
15	230	126	52.900	15.876	28.980
16	235	122	55.225	14.884	28.670
17	235	122	55.225	14.884	28.670
18	235	122	55.225	14.884	28.670
19	235	119	55.225	14.161	27.965
20	235	122	55.225	14.884	28.670
Soma	4.550	2.588	1.035.750	335.594	588.125
Média	227,5	129,4			

Soma de Quadrados:

$$\begin{aligned} \square S_{xx} &= \sum_{i=1}^n x_i^2 - n\bar{x}^2 = 1035750 - 20 * 227,50^2 = \\ &625 \\ \square S_{yy} &= \sum_{i=1}^n y_i^2 - n\bar{y}^2 = 335594 - 20 * 129,40^2 = \\ &706,80 \\ \square S_{xy} &= \sum_{i=1}^n x_i y_i - n\bar{x}\bar{y} = 588125 - 20 * 227,5 * \\ &129,4 = -645 \end{aligned}$$

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{-645}{625} = -1,032$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 129,4 - (-1,032) * 227,5 = 364,18$$

O modelo ajustado é dado por:

$$\text{Dureza} = 364,18 - 1,032 \text{ Temperatura}$$

Propriedade dos estimadores

Resumo



$$E(\hat{\beta}_1) = \beta_1 \text{ e } Var(\hat{\beta}_1) = \frac{\sigma^2}{S_{xx}}$$



$$E(\hat{\beta}_0) = \beta_0 \text{ e } Var(\hat{\beta}_0) = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right)$$



$$Cov(\hat{\beta}_0, \hat{\beta}_1) = -\frac{\bar{x}\sigma^2}{S_{xx}}$$



$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{S_{xx}}\right)$$



$$\hat{\beta}_0 \sim N\left(\beta_0, \frac{\sigma^2}{n} + \frac{\sigma^2 \bar{x}^2}{S_{xx}}\right)$$



$$E(QME) = \frac{E(SQE)}{n-2} = \frac{E(\sum_{i=1}^n e_i^2)}{n-2} = \frac{E(\sum_{i=1}^n (y_i - \hat{y}_i)^2)}{n-2} = \sigma^2$$

$$SQE = S_{yy} - \hat{\beta}_1 S_{xy}$$

$$\sum_{i=1}^n e_i = \sum_{i=1}^n (y_i - \hat{y}_i) = 0$$

Exemplo

Em problemas de tratamento térmico deseja-se estabelecer uma relação entre a temperatura da estufa e uma característica da qualidade (Dureza). Determinar os valores de temperatura em C que “otimizam” a performance do processo de tratamento térmico em relação a estrutura metalográfica do material, avaliada em relação de dureza em HB.

Observação	Temperatura (x)	Dureza (y)	x^2	y^2	$x \times y$
1	220	137	48.400	18.769	30.140
2	220	137	48.400	18.769	30.140
3	220	137	48.400	18.769	30.140
4	220	136	48.400	18.496	29.920
5	220	135	48.400	18.225	29.700
6	225	135	50.625	18.225	30.375
7	225	133	50.625	17.689	29.925
8	225	132	50.625	17.424	29.700
9	225	133	50.625	17.689	29.925
10	225	133	50.625	17.689	29.925
11	230	128	52.900	16.384	29.440
12	230	124	52.900	15.376	28.520
13	230	126	52.900	15.876	28.980
14	230	129	52.900	16.641	29.670
15	230	126	52.900	15.876	28.980
16	235	122	55.225	14.884	28.670
17	235	122	55.225	14.884	28.670
18	235	122	55.225	14.884	28.670
19	235	119	55.225	14.161	27.965
20	235	122	55.225	14.884	28.670
Soma	4.550	2.588	1.035.750	335.594	588.125
Média	227,5	129,4			

Temos que:

$$\begin{aligned} \square S_{xx} &= \sum_{i=1}^n x_i^2 - n\bar{x}^2 = 1035750 - 20 * 227,50^2 = \\ &= 625 \\ \square S_{yy} &= \sum_{i=1}^n y_i^2 - n\bar{y}^2 = 335594 - 20 * 129,40^2 = \\ &= 706,80 \\ \square S_{xy} &= \sum_{i=1}^n x_i y_i - n\bar{x}\bar{y} = 588125 - 20 * 227,5 * \\ &129,4 = -645 \\ \square \hat{\beta}_1 &= \frac{S_{xy}}{S_{xx}} = \frac{-645}{625} = -1,032 \end{aligned}$$

$$SQE = S_{yy} - \hat{\beta}_1 S_{xy} = 706,8 - (-1,032) * (-645) \\ = 41,16$$

$$\hat{\sigma}^2 = QME = \frac{SQE}{n - 2} = \frac{41,16}{20 - 2} = 2,2866$$

$$\widehat{Var}(\hat{\beta}_0) = QME \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right) = 2,2866 \left(\frac{1}{20} + \frac{227,5^2}{625} \right) = 189,47$$

$$\widehat{Var}(\hat{\beta}_1) = \frac{QME}{S_{xx}} = \frac{2,2866}{625} = 0,003658$$

Inferência para o parâmetro β_0

Teste para o Coeficiente linear (Intercepto):

$$\begin{cases} H_0: \beta_0 = 0 \\ H_1: \beta_0 \neq 0 \end{cases}$$

no qual

$$QME = \frac{S_{yy} - \hat{b}S_{xy}}{n - 2} = \frac{SQE}{n - 2}$$

Estatística do Teste

$$T_0 = \frac{\hat{\beta}_0}{\sqrt{\widehat{Var}(\hat{\beta}_0)}} \sim t_{n-2} \text{ (Sob } H_0)$$

$$\widehat{Var}(\hat{\beta}_0) = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right) = QME \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right)$$

O valor de T_0 deve ser comparado com o quantil distribuição t- Student com $n-2$ graus de liberdade para um determinado nível de significância α

- Se $|T_0| \leq t_{(n-2; 1-\frac{\alpha}{2})}$ não rejeitamos H_0 , ou seja, rejeitamos a hipótese de que o coeficiente angular seja significativo;
- Se $|T_0| > t_{(n-2; 1-\frac{\alpha}{2})}$ rejeitamos H_0 , ou seja, não rejeitamos a hipótese de que o coeficiente angular seja significativo;

- Outra forma de definirmos um critério para avaliarmos o teste de hipótese é o P-valor.
- O P-valor = $2 * P(t_{(n-2)} > |T_0|)$ representa o menor nível de significância para o qual rejeitamos H_0 . Logo, para um nível de significância $\alpha = 0,05$ adotado (por exemplo), rejeitamos H_0 se o P-valor obtido for menor que 0,05, enquanto que não rejeitamos H_0 se o P-valor for maior que 0,05.

Intervalo de Confiança para β_0 com $(1 - \alpha)100\%$ é dado por:

$$\left[\hat{\beta}_0 - t_{(1-\frac{\alpha}{2}; n-2)} \sqrt{QME \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right)} ; \hat{\beta}_0 + t_{(1-\frac{\alpha}{2}; n-2)} \sqrt{QME \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right)} \right]$$

Inferência para o parâmetro β_1

Teste para o Coeficiente Angular:

$$\begin{cases} H_0: \beta_1 = 0 \\ H_1: \beta_1 \neq 0 \end{cases}$$

Estatística do Teste

$$T_0 = \frac{\hat{\beta}_1}{\sqrt{\widehat{Var}(\hat{\beta}_1)}} \sim t_{n-2} \text{ (Sob } H_0)$$

O valor de T_0 deve ser comparado com o quantil distribuição t- Student com $n-2$ graus de liberdade para um determinado nível de significância α

- Se $|T_0| \leq t_{(n-2; 1-\frac{\alpha}{2})}$ não rejeitamos H_0 , ou seja, rejeitamos a hipótese de que o coeficiente angular seja significativo;
- Se $|T_0| > t_{(n-2; 1-\frac{\alpha}{2})}$ rejeitamos H_0 , ou seja, não rejeitamos a hipótese de que o coeficiente angular seja significativo;

- Outra forma de definirmos um critério para avaliarmos o teste de hipótese é o P-valor.
- O P-valor = $2 * P(t_{(n-2)} > |T_0|)$ representa o menor nível de significância para o qual rejeitamos H_0 . Logo, para um nível de significância $\alpha = 0,05$ adotado (por exemplo), rejeitamos H_0 se o P-valor obtido for menor que 0,05, enquanto que não rejeitamos H_0 se o P-valor for maior que 0,05.

Intervalo de Confiança para β_1 com $(1 - \alpha)100\%$ é dado por:

$$\left[\hat{\beta}_1 - t_{(1-\frac{\alpha}{2}; n-2)} \sqrt{\frac{QME}{S_{xx}}} ; \hat{\beta}_1 + t_{(1-\frac{\alpha}{2}; n-2)} \sqrt{\frac{QME}{S_{xx}}} \right]$$

no qual

$$QME = \frac{S_{yy} - \hat{b}S_{xy}}{n-2} = \frac{SQE}{n-2}$$

$$\widehat{Var}(\hat{\beta}_1) = \frac{\sigma^2}{S_{xx}} = \frac{QME}{S_{xx}}$$

Exemplo

Do exemplo anterior, obtemos que:

$$\widehat{Var}(\hat{\beta}_0) = QME \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right) = 2,2866 \left(\frac{1}{20} + \frac{227,5^2}{625} \right) = 189,47$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 129,4 - (-1,032) * 227,5 = 364,18$$

Em problemas de tratamento térmico deseja-se estabelecer uma relação entre a temperatura da estufa e uma característica da qualidade (Dureza). Determinar os valores de temperatura em C que “otimizam” a performance do processo de tratamento térmico em relação a estrutura metalográfica do material, avaliada em relação de dureza em HB.

Assim, temos o Teste para o Coeficiente linear (Intercepto):

$$\begin{cases} H_0: \beta_0 = 0 \\ H_1: \beta_0 \neq 0 \end{cases}$$

Estatística do Teste

$$T_0 = \frac{\hat{\beta}_0}{\sqrt{\widehat{Var}(\hat{\beta}_0)}} = \frac{364,18}{\sqrt{189,47}} = 26,46$$

O valor de $|T_0|$ deve ser comparado com $t_{(0,95; 18)} = 2,101$ (distribuição t- Student com 18 graus de liberdade) para um nível de significância $\alpha = 0,05$.

Desde que $|T_0| = 26,46 > 2,101$ ou P-valor = $2 * P(t_{(n-2)} > |T_0|) \cong 0$, rejeitamos a hipótese H_0

Intervalo de Confiança para β_1 com $(1 - 0,05)100\%$ é dado por:

$$\left[\hat{\beta}_0 - t_{\left(1 - \frac{0,05}{2}; 20-2\right)} \sqrt{\frac{QME}{S_{xx}}} ; \hat{\beta}_0 + t_{\left(1 - \frac{0,05}{2}; 20-2\right)} \sqrt{\frac{QME}{S_{xx}}} \right] = [335,26 ; 393,09]$$

Exemplo

Do exemplo anterior, obtemos que:

$$\widehat{Var}(\hat{\beta}_1) = \frac{QME}{S_{xx}} = \frac{2,2866}{625} = 0,003658$$

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{-645}{625} = -1,032$$

Em problemas de tratamento térmico deseja-se estabelecer uma relação entre a temperatura da estufa e uma característica da qualidade (Dureza). Determinar os valores de temperatura em C que “otimizam” a performance do processo de tratamento térmico em relação a estrutura metalográfica do material, avaliada em relação de dureza em HB.

Assim, temos o Teste para o Coeficiente Angular:

$$\begin{cases} H_0: \beta_1 = 0 \\ H_1: \beta_1 \neq 0 \end{cases}$$

Estatística do Teste

$$T_0 = \frac{\hat{\beta}_1}{\sqrt{\widehat{Var}(\hat{\beta}_1)}} = \frac{-1,032}{\sqrt{0,003658}} = -17,06$$

O valor de $|T_0|$ deve ser comparado com $t_{(0,95; 18)} = 2,101$ (distribuição t- Student com 18 graus de liberdade) para um nível de significância $\alpha = 0,05$.

Desde que $|T_0| = 17,06 > 2,101$ ou P-valor = $2 * P(t_{(n-2)} > |T_0|) \cong 0$ rejeitamos a hipótese H_0

Intervalo de Confiança para β_1 com $(1 - 0,05)100\%$ é dado por:

$$\left[\hat{\beta}_1 - t_{(1-\frac{\alpha}{2}; n-2)} \sqrt{\frac{QME}{S_{xx}}} ; \hat{\beta}_1 + t_{(1-\frac{\alpha}{2}; n-2)} \sqrt{\frac{QME}{S_{xx}}} \right] = [-1,159 ; -0,905]$$

ANOVA

Análise de Variância

Análise de Variância

Soma de Quadrados

$$SQT = SQR + SQE$$

$$\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

Dividindo-se a soma de quadrados pelos respectivos graus de liberdade, obtemos o quadrado médio

Quadrados Médios

$$\text{Quadrado Médio da Regressão: } QMR = \frac{SQR}{1} = SQR$$

$$\text{Quadrado Médio do Erro (Resíduo): } QME = \frac{SQE}{n-2} = \frac{S_{yy} - \hat{\beta}_1 S_{xy}}{n-2}$$

ANOVA

Fonte	GL	Soma de quadrados	Quadrado médio
Regressão	1	$SQR = \hat{\beta}_1 S_{xy}$	$QMR = SQR$
Resíduo	$n - 2$	$SQE = S_{yy} - \hat{\beta}_1 S_{xy}$	$QME = \frac{SQE}{(n - 2)}$
Total	$n - 1$	S_{yy}	

Análise de Variância

Considerando o Modelo de Regressão Linear Simples, a análise de regressão estabelece um teste para avaliar o parâmetro β_1

Sob H_0 , temos que:

$$\begin{cases} H_0: \beta_1 = 0 \\ H_1: \beta_1 \neq 0 \end{cases}$$

$$\frac{1}{\sigma^2} \sum_{i=1}^n (y_i - \bar{y})^2 = \frac{1}{\sigma^2} \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + \frac{1}{\sigma^2} \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

Para estabelecer a independência entre as SQE e a SQA, vamos utilizar o teorema de Cochran.

Teorema de Cochran

Sejam Z_1, Z_2, \dots, Z_p variáveis aleatórias independentes com distribuição $N(0, 1)$. Então

$$\sum_{i=1}^p Z_i^2 \text{ possui distribuição } \chi_{(p)}^2$$

Então, pelo teorema de Cochran temos,

$$\chi_T = \frac{n-1}{\sigma^2} \sum_{i=1}^n (y_i - \bar{y})^2 \sim \chi_{(n-1)}^2$$

$$\chi_E = \frac{n-2}{\sigma^2} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \sim \chi_{(n-2)}^2$$

$$\chi_R = \frac{1}{\sigma^2} \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 \sim \chi_{(1)}^2$$

As distribuições Qui-quadradas χ_E e χ_R são independentes

Análise de Variância

Estatística do Teste

$$F_0 = \frac{\frac{\chi_R}{1}}{\frac{\chi_E}{n-2}} = \frac{QMR}{QME} \sim F_{(1,n-2)}$$

O valor de F_0 deve ser comparado com o quantil distribuição F de Snedecor com 1 grau de liberdade no numerador e $n-2$ graus de liberdade no denominador para um determinado nível de significância α

- Se $F_0 \leq F_{(1;n-2;1-\alpha)}$ não rejeitamos H_0 , ou seja, rejeitamos a hipótese de que o coeficiente angular seja significativo;
- Se $F_0 > F_{(1;n-2;1-\alpha)}$ rejeitamos H_0 , ou seja, não rejeitamos a hipótese de que o coeficiente angular seja significativo;

- Outra forma de definirmos um critério para avaliarmos o teste de hipótese é o P-valor.
- O P-valor = $P(F_{(1;n-2)} > F_0)$ representa o menor nível de significância para o qual rejeitamos H_0 . Logo, para um nível de significância $\alpha = 0,05$ adotado (por exemplo), rejeitamos H_0 se o P-valor obtido for menor que 0,05, enquanto que não rejeitamos H_0 se o P-valor for maior que 0,05.

Fonte	GL	Soma de quadrados	Quadrado médio
Regressão	1	$SQR = \hat{\beta}_1 S_{xy}$	$QMR = SQR$
Resíduo	$n - 2$	$SQE = S_{yy} - \hat{\beta}_1 S_{xy}$	$QME = \frac{SQE}{(n - 2)}$
Total	$n - 1$	S_{yy}	

Exemplo

Do exemplo anterior, obtemos que:

$$SQT = S_{yy} = 706,8 \text{ e } SQE = S_{yy} - \hat{\beta}_1 S_{xy} = 41,16$$

$$SQR = SQT - SQE = 706,8 - 41,16 = 665,64$$

Em problemas de tratamento térmico deseja-se estabelecer uma relação entre a temperatura da estufa e uma característica da qualidade (Dureza). Determinar os valores de temperatura em C que “otimizam” a performance do processo de tratamento térmico em relação a estrutura metalográfica do material, avaliada em relação de dureza em HB.

Assim, temos o Teste para avaliar β_1 :

$$\begin{cases} H_0: \beta_1 = 0 \\ H_1: \beta_1 \neq 0 \end{cases}$$

Estatística do Teste

$$F_0 = \frac{QMR}{QME} = \frac{\frac{665,64}{1}}{\frac{41,16}{18}} = 291,10$$

O valor de F_0 deve ser comparado com $F_{(0,95; 1; 18)} = 4,4138$ (distribuição F com 1 grau de liberdade no numerador e 18 graus de liberdade no denominador) para um nível de significância $\alpha = 0,05$.

Desde que $F_0 = 291,1 > 4,4138$ ou P-valor = $P(F_{(1;20-2)} > F_0) \cong 0$ rejeitamos a hipótese H_0

Fonte	GL	Soma de quadrados	Quadrado médio	F_0
Regressão	1	665,64	$\frac{665,64}{1} = 665,64$	$\frac{665,64}{2,29} = 291,10$
Resíduo	18	41,16	$\frac{41,16}{18} = 2,29$	
Total	19	706,80		

Medidas de Associação

Medidas de Associação

Coeficiente de determinação R^2



$$R^2 = \frac{SQR}{Syy} = 1 - \frac{SQE}{Syy} = \frac{\hat{\beta}_1 S_{xy}}{S_{yy}} \text{ com } 0 \leq R^2 \leq 1$$



É uma medida descritiva da qualidade do ajuste obtido, em geral, referimos como a quantidade de variabilidade nos dados que é explicada pelo modelo de regressão ajustado



o valor do coeficiente de determinação depende do número de observações (n), tendendo a crescer quando n diminui. Se $n = 2$ temos sempre $R^2 = 1$

Precauções no uso R^2



é sempre possível torná-lo maior pela adição de um número suficiente de termos ao modelo. Por exemplo, se não há dados repetidos (mais do que um valor para um mesmo x) um polinômio de grau $(n - 1)$ dará um ajuste perfeito ($R^2 = 1$) para n dados.



Embora (R^2) aumente com a adição de termos ao modelo, isto não significa necessariamente que o novo modelo é superior ao anterior. A menos que a soma de quadrados residual do novo modelo seja reduzida por uma quantidade igual ao quadrado médio residual original, o novo modelo terá um quadrado médio residual maior do que o original, devido a perda de 1 grau de liberdade.

Exemplo

Em problemas de tratamento térmico deseja-se estabelecer uma relação entre a temperatura da estufa e uma característica da qualidade (Dureza). Determinar os valores de temperatura em C que “otimizam” a performance do processo de tratamento térmico em relação a estrutura metalográfica do material, avaliada em relação de dureza em HB.

Observação	Temperatura (x)	Dureza (y)	x^2	y^2	$x \times y$
1	220	137	48.400	18.769	30.140
2	220	137	48.400	18.769	30.140
3	220	137	48.400	18.769	30.140
4	220	136	48.400	18.496	29.920
5	220	135	48.400	18.225	29.700
6	225	135	50.625	18.225	30.375
7	225	133	50.625	17.689	29.925
8	225	132	50.625	17.424	29.700
9	225	133	50.625	17.689	29.925
10	225	133	50.625	17.689	29.925
11	230	128	52.900	16.384	29.440
12	230	124	52.900	15.376	28.520
13	230	126	52.900	15.876	28.980
14	230	129	52.900	16.641	29.670
15	230	126	52.900	15.876	28.980
16	235	122	55.225	14.884	28.670
17	235	122	55.225	14.884	28.670
18	235	122	55.225	14.884	28.670
19	235	119	55.225	14.161	27.965
20	235	122	55.225	14.884	28.670
Soma	4.550	2.588	1.035.750	335.594	588.125
Média	227,5	129,4			

Temos que:

$$\begin{aligned} \square S_{xx} &= \sum_{i=1}^n x_i^2 - n\bar{x}^2 = 1035750 - 20 * 227,50^2 = \\ &625 \\ \square S_{yy} &= \sum_{i=1}^n y_i^2 - n\bar{y}^2 = 335594 - 20 * 129,40^2 = \\ &706,80 \\ \square S_{xy} &= \sum_{i=1}^n x_i y_i - n\bar{x}\bar{y} = 588125 - 20 * 227,5 * \\ &129,4 = -645 \\ \square \hat{\beta}_1 &= \frac{S_{xy}}{S_{xx}} = \frac{-645}{625} = -1,032 \end{aligned}$$

$$R^2 = \frac{\hat{\beta}_1 S_{xy}}{S_{yy}} = \frac{-1,032 * (-645)}{706,8} = 0,9417$$

Intervalo de Confiança para média

Intervalo de Confiança para Média

Intervalo de Confiança para $E(Y|X = x_0) = \mu_{y|x_0} = \beta_0 + \beta_1 x_0$

Estimador Pontual

$$\hat{y} = \hat{\mu}_{y|x_0} = \hat{\beta}_0 + \hat{\beta}_1 x_0$$

Considerando, $\varepsilon_i \sim^{iid} N(0, \sigma^2)$, temos:

$$E(\hat{\mu}_{y|x_0}) = \mu_{y|x_0} = \beta_0 + \beta_1 x_0 = E(y(x_0)) \text{ e}$$
$$Var(\hat{\mu}_{y|x_0}) = \sigma^2 \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right)$$

Assim,

$$t = \frac{\hat{y}(x_0) - \mu_{y|x_0}}{\sqrt{QME \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right)}} \sim t_{(n-2)}$$

Intervalo de Confiança para $\mu_{y|x_0}$ com $(1 - \alpha)100\%$ é dado por:

$$\left[\hat{y} - t_{(1-\frac{\alpha}{2}; n-2)} \sqrt{QME \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right)} ; \hat{y} + t_{(1-\frac{\alpha}{2}; n-2)} \sqrt{QME \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right)} \right]$$

Exemplo

Em problemas de tratamento térmico deseja-se estabelecer uma relação entre a temperatura da estufa e uma característica da qualidade (Dureza). Determinar os valores de temperatura em C que “otimizam” a performance do processo de tratamento térmico em relação a estrutura metalográfica do material, avaliada em relação de dureza em HB.

Do exemplo anterior, obtemos que:

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{-645}{625} = -1,032 \text{ e } SQE = S_{yy} - \hat{\beta}_1 S_{xy} = 41,16$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 129,4 - (-1,032) * 227,5 = 364,18$$

Para $x_0 = 220$:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_0 = 364,18 - 1,032 * 220 = 137,14$$

Soma de Quadrados:

- $S_{xx} = \sum_{i=1}^n x_i^2 - n\bar{x}^2 = 1035750 - 20 * 227,50^2 = 625$
- $S_{yy} = \sum_{i=1}^n y_i^2 - n\bar{y}^2 = 335594 - 20 * 129,40^2 = 706,80$
- $S_{xy} = \sum_{i=1}^n x_i y_i - n\bar{x}\bar{y} = 588125 - 20 * 227,5 * 129,4 = -645$

$$\sqrt{QME \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right)} = \sqrt{\frac{41,16}{20-2} \left(\frac{1}{20} + \frac{(220 - 227,5)^2}{625} \right)} = 0,5628$$

Intervalo de Confiança para $\mu_{y|x_0}$ com $(1 - \alpha)100\%$ é dado por:

$$\left[\hat{y} - t_{(1-\frac{\alpha}{2};n-2)} \sqrt{QME \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right)} ; \hat{y} + t_{(1-\frac{\alpha}{2};n-2)} \sqrt{QME \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right)} \right] = [135,95 ; 138,33]$$

Intervalo de Predição

Intervalo de Predição

Intervalo de Predição

Estimador Pontual

$$\hat{y}_h = \hat{\beta}_0 + \hat{\beta}_1 x_h$$

Chamamos de erro na previsão a diferença $(y_h - \hat{y}_h)$, cuja variância é dada por:

$$Var(\hat{\mu}_{y|x_0}) = \sigma^2 \left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right)$$

Assim,

$$t = \frac{y_h - \hat{y}_h}{\sqrt{QME \left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right)}} \sim t_{(n-2)}$$

Intervalo de Confiança para y_h com $(1 - \alpha)100\%$ é dado por:

$$\left[\hat{y}_h - t_{(1-\frac{\alpha}{2}; n-2)} \sqrt{QME \left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right)} ; \hat{y}_h + t_{(1-\frac{\alpha}{2}; n-2)} \sqrt{QME \left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right)} \right]$$

Exemplo

Em problemas de tratamento térmico deseja-se estabelecer uma relação entre a temperatura da estufa e uma característica da qualidade (Dureza). Determinar os valores de temperatura em C que “otimizam” a performance do processo de tratamento térmico em relação a estrutura metalográfica do material, avaliada em relação de dureza em HB.

Do exemplo anterior, obtemos que:

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{-645}{625} = -1,032 \text{ e } SQE = S_{yy} - \hat{\beta}_1 S_{xy} = 41,16$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 129,4 - (-1,032) * 227,5 = 364,18$$

Para $x_h = 217,5$:

$$\hat{y}_h = \hat{\beta}_0 + \hat{\beta}_1 x_h = 364,18 - 1,032 * 217,5 = 139,72$$

Soma de Quadrados:

- $S_{xx} = \sum_{i=1}^n x_i^2 - n\bar{x}^2 = 1035750 - 20 * 227,50^2 = 625$
- $S_{yy} = \sum_{i=1}^n y_i^2 - n\bar{y}^2 = 335594 - 20 * 129,40^2 = 706,80$
- $S_{xy} = \sum_{i=1}^n x_i y_i - n\bar{x}\bar{y} = 588125 - 20 * 227,5 * 129,4 = -645$

$$\sqrt{QME \left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right)} = \sqrt{\frac{41,16}{20-2} \left(1 + \frac{1}{20} + \frac{(220 - 227,5)^2}{625} \right)} = 1,6634$$

Intervalo de Confiança para y_h com $(1 - \alpha)100\%$ é dado por:

$$\left[\hat{y}_h - t_{\left(1-\frac{\alpha}{2}; n-2\right)} \sqrt{QME \left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right)} ; \hat{y}_h + t_{\left(1-\frac{\alpha}{2}; n-2\right)} \sqrt{QME \left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right)} \right] \\ = [136,22 ; 143,21]$$

Modelo de Regressão sem Intercepto

Modelo Estatístico

Definição

Dados n pares de valores, $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$, se y é função linear de x , pode-se estabelecer uma regressão linear simples sem intercepto, cujo modelo estatístico

$$Y_i = \beta_1 x_i + \varepsilon_i$$

para $i = 1, \dots, n.$

Neste modelo:

- Y_i é uma variável aleatória e representa o valor da variável resposta;
- x_i são as variáveis explicativas;
- ε_i uma variável aleatória que representa o erro experimental;
- β_1 é o parâmetro do modelo, que será estimado, e que define a reta de regressão;
- n é o tamanho da amostra.

Estimação dos parâmetros

Estimativas
de Mínimos
Quadrados

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2}$$

$$E(\hat{\beta}_1) = \beta_1$$

$$Var(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^n x_i^2}$$

A reta de regressão ajustada é dada por

$$\hat{y}_i = \hat{\beta}_1 x_i ,
i = 1, 2, \dots, n.$$

Propriedade dos estimadores

Resumo



$$E(\hat{\beta}_1) = \beta_1 \text{ e } Var(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^n x_i^2}$$



$$\varepsilon_i \stackrel{iid}{\sim} N(0, \sigma^2)$$



$$\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{\sigma^2}{\sum_{i=1}^n x_i^2}}} \sim N(0,1)$$



$$E(QME) = \frac{E(SQE)}{n-1} = \frac{E\left(\sum_{i=1}^n e_i^2\right)}{n-1} = \frac{E\left(\sum_{i=1}^n (y_i - \hat{y}_i)^2\right)}{n-1} = \sigma^2$$

$$\sum_{i=1}^n e_i = \sum_{i=1}^n (y_i - \hat{y}_i) = 0$$

Modelo de regressão sem intercepto

Estatística do Teste

$$T_0 = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{\sigma^2}{\sum_{i=1}^n x_i^2}}} \sim t_{n-1} \text{ (Sob } H_0)$$

Intervalo de Confiança para β_1 com $(1 - \alpha)100\%$ é dado por:

$$\left[\hat{\beta}_1 - t_{(\frac{\alpha}{2}; n-1)} \sqrt{\frac{QME}{\sum_{i=1}^n x_i^2}} ; \hat{\beta}_1 + t_{(\frac{\alpha}{2}; n-1)} \sqrt{\frac{QME}{\sum_{i=1}^n x_i^2}} \right]$$

Intervalo de Confiança para a resposta média $X = x_0$ com $(1 - \alpha)100\%$ é dado por:

$$\left[\hat{\mu}_{y|x_0} - t_{(\frac{\alpha}{2}; n-1)} \sqrt{\frac{QME}{\sum_{i=1}^n x_i^2}} ; \hat{\mu}_{y|x_0} + t_{(\frac{\alpha}{2}; n-1)} \sqrt{\frac{QME}{\sum_{i=1}^n x_i^2}} \right]$$

Em que
 $\hat{\mu}_{y|x_0} = \hat{\beta}_1 x_0$

Intervalo de previsão para $X = x_h$ com $(1 - \alpha)100\%$ é dado por:

$$\left[\hat{y}_h - t_{(\frac{\alpha}{2}; n-1)} \sqrt{QME \left(1 + \frac{x_0}{\sum_{i=1}^n x_i^2} \right)} ; \hat{y}_h + t_{(\frac{\alpha}{2}; n-1)} \sqrt{QME \left(1 + \frac{x_0}{\sum_{i=1}^n x_i^2} \right)} \right]$$

Em que
 $\hat{y}_h = \hat{\beta}_1 x_h$

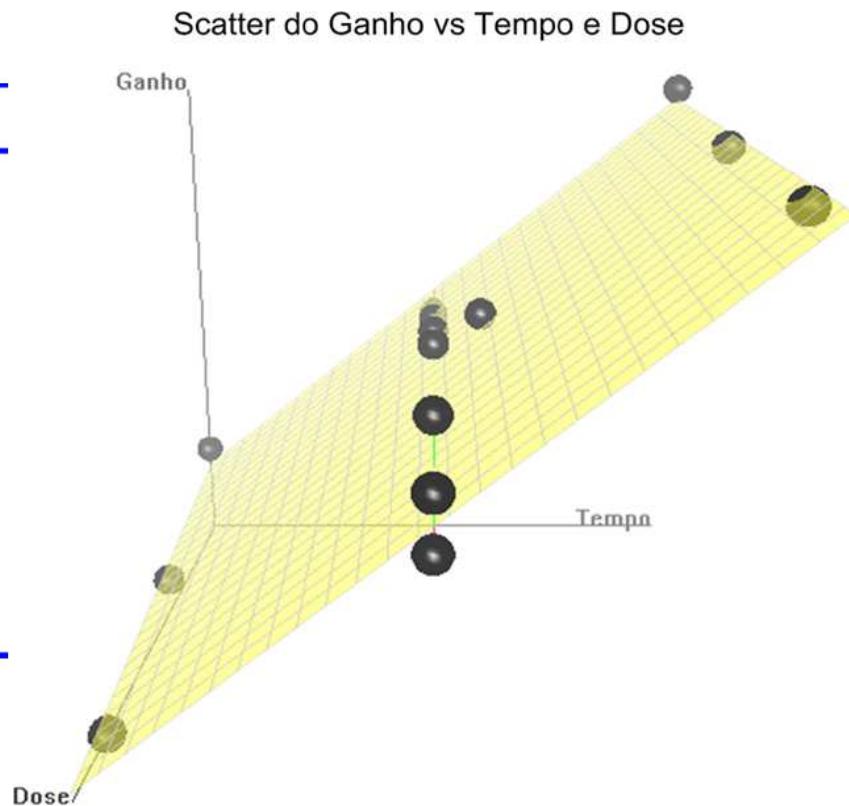
Análise de Regressão

Regressão Linear Múltipla

Motivação 2

O ganho de um transistor consiste na diferença entre o emissor e o coletor. A variável **Ganho** (em hFE) pode ser controlada no processo de deposição de íons por meio das variáveis **Tempo de emissão** (em minutos) e **Dose de íons** ($\times 10^{14}$).

Observação	Tempo (min)	Dose de íons ($\times 10^{14}$)	Ganho (hFE)
1	195	4	1004
2	255	4	1636
3	195	4,6	852
4	255	4,6	1506
5	225	4,2	1272
6	225	4,1	1270
7	225	4,6	1269
8	195	4,3	903
9	255	4,3	1555
10	225	4	1260
11	225	4,7	1146
12	225	4,3	1276
13	225	4,72	1225
14	230	4,3	1321



Regressão Linear Múltipla

Objetivos

- Modelo Estatístico
- Estimação dos parâmetros
- Propriedades dos Estimadores
- Testes de hipóteses e intervalos de confiança para os parâmetros do modelo
- Análise de diagnóstico do modelo
- Aplicações

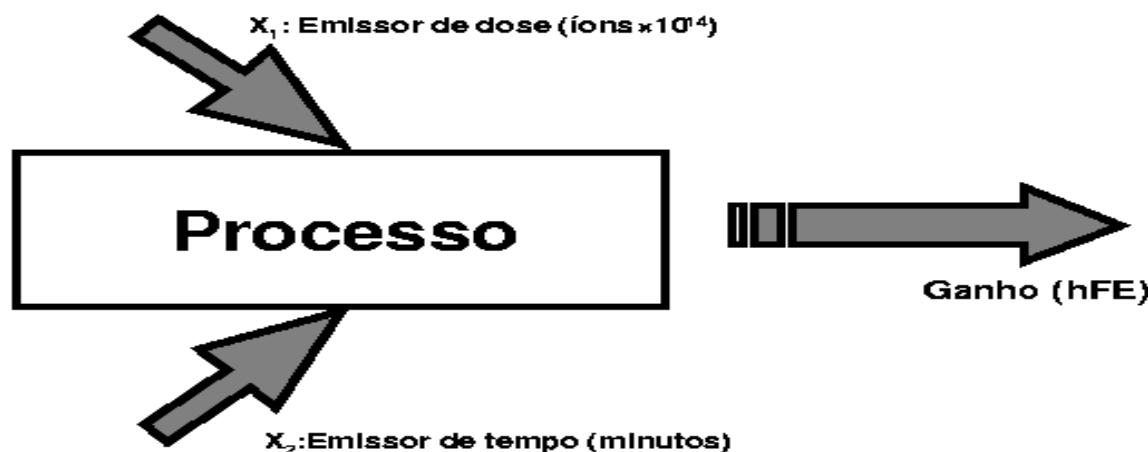
Modelo Estatístico

Um modelo de regressão múltipla, com duas variáveis explicativas:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

Em que:

- β_0 corresponde ao intercepto do plano com o eixo z. Se $x = (x_1; x_2) = (0, 0)$ o parâmetro β_0 fornece a resposta média nesse ponto. Caso contrário, não é possível interpretar o parâmetro β_0 .
- β_1 indica uma mudança na resposta média a cada unidade de mudança em x_1 , quando as demais variáveis são mantidas fixas.
- β_2 indica uma mudança na resposta média a cada unidade de mudança em x_2 , quando x_1 é mantido constante.



Exemplo

O ganho de um transistor consiste na diferença entre o emissor e o coletor. A variável **Ganho** (em hFE) pode ser controlada no processo de deposição de íons por meio das variáveis **Tempo de emissão** (em minutos) e **Dose de íons** ($\times 10^{14}$).

Observação	Tempo (min)	Dose de íons ($\times 10^{14}$)	Ganho (hFE)
1	195	4	1004
2	255	4	1636
3	195	4,6	852
4	255	4,6	1506
5	225	4,2	1272
6	225	4,1	1270
7	225	4,6	1269
8	195	4,3	903
9	255	4,3	1555
10	225	4	1260
11	225	4,7	1146
12	225	4,3	1276
13	225	4,72	1225
14	230	4,3	1321

Para esses dados temos o modelo:

$$E(Y|x) = -520,08 - 152,15 \text{ Dose} + 10,78 \text{ Tempo}$$

Se o Tempo = 225, (constante) então

$$E(Y|x) = 1905,42 - 152,15 \text{ Dose}$$

Assim:

- $\beta_1 = -152,15$ indica que a cada acréscimo de uma unidade na Dose a resposta média decrescerá 152,15 unidades
- $\beta_2 = 10,78$ indica um acréscimo na resposta média de 10,78 unidades para cada acréscimo de uma unidade na variável Dose.

Efeito das Interações

Seja o modelo

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 \underbrace{x_1 x_2}_{\text{interação}} + \varepsilon$$

Comentários:

- Se a interação está presente e é significativa, o efeito de x_1 na resposta média depende do nível de x_2 e analogamente o efeito de x_2 na resposta média depende do nível de x_1
- A interpretação para os parâmetros β_1 e β_2 , no modelo com interação, não é o mesmo visto anteriormente.

Suposições do Modelo

Modelo de Regressão Linear Múltipla (MRLM)

- O erro tem média zero e variância σ^2 , desconhecida
- Os erros são não correlacionados
- Os erros têm distribuição normal
- As variáveis regressoras x_1, x_2, \dots, x_p assumem valores fixos

Se as suposições do MRLM se verificam, então a variável Y tem distribuição normal com variância σ^2 e média

$$E(Y|x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

Estimação dos parâmetros do modelo

Os dados de um MRLM podem ser representados da seguinte forma:

y	x ₁	x ₂	...	x _p
y ₁	x ₁₁	x ₁₁	...	x _{1p}
y ₂	x ₂₂	x ₂₂	...	x _{2p}
⋮	⋮	⋮	⋮	⋮
y _n	x _{n1}	x _{n2}	...	x _{np}

Cada observação satisfaz:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i, \\ i = 1, \dots, n$$

Estimação dos parâmetros

Método dos Mínimos Quadrados

Estimar os parâmetros $\beta_0, \beta_1, \dots, \beta_p$ de modo que os desvios
 $\varepsilon_i = Y_i - [\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip}]$
sejam mínimos

Consiste em minimizar L, que é a soma dos quadrados dos desvios

$$L = \sum_{i=1}^n \varepsilon_i^2 = \sum_{i=1}^n [Y_i - (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip})]^2$$

Lembrando que $\sum_{i=1}^n \varepsilon_i = 0$

Passos para minimizar L

Derivar a função L, em relação aos parâmetros

Substituir $\beta_0, \beta_1, \dots, \beta_p$ por $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p$

Igualar as derivadas à zero

Resolver o sistema de equações, em relação as variáveis
 $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p$

Estimação dos parâmetros

Representação
Matricial



$$Y = X\beta + \varepsilon$$

Em que

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad X = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1p} \\ 1 & x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{np} \end{bmatrix}, \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} \text{ e } \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}.$$

O método de mínimos quadrados tem como objetivo encontrar o vetor $\hat{\beta}$ que minimiza:

$$L = \sum_{i=1}^n \varepsilon_i^2 = \varepsilon' \varepsilon = Y'Y - 2\beta'X'Y + \beta'X'X\beta.$$

$$\hat{\beta} = (X'X)^{-1}X'Y.$$

Portanto, o modelo de regressão linear ajustado e o vetor de resíduos, são respectivamente:

$$\hat{Y} = X\hat{\beta} \quad \text{e} \quad e = Y - \hat{Y}.$$

Exemplo

O ganho de um transistor consiste na diferença entre o emissor e o coletor. A variável **Ganho** (em hFE) pode ser controlada no processo de deposição de íons por meio das variáveis **Tempo de emissão (em minutos)** e **Dose de íons ($\times 10^{14}$)**.

Observação	Tempo (min)	Dose de íons ($\times 10^{14}$)	Ganho (hFE)
1	195	4	1004
2	255	4	1636
3	195	4,6	852
4	255	4,6	1506
5	225	4,2	1272
6	225	4,1	1270
7	225	4,6	1269
8	195	4,3	903
9	255	4,3	1555
10	225	4	1260
11	225	4,7	1146
12	225	4,3	1276
13	225	4,72	1225
14	230	4,3	1321

Sejam: Ganho (Y_i), Dose (x_1) e Tempo (x_2).

$$n = 14, \quad \sum_{i=1}^{14} Y_i = 17.495, \quad \sum_{i=1}^{14} x_1 = 60,72 \quad \text{e} \quad \sum_{i=1}^{14} x_2 = 3.155$$

$$\sum_{i=1}^{14} x_{i1}^2 = 264,26, \quad \sum_{i=1}^{14} x_{i2}^2 = 716425 \quad \text{e} \quad \sum_{i=1}^{14} x_{i1}x_{i2} = 13.683,50$$

$$\sum_{i=1}^{14} x_{i1}Y_i = 75.738,30, \quad \sum_{i=1}^{14} x_{i2}Y_i = 4.001.120.$$

Substituindo os valores para esse exemplo, temos:

$$\begin{cases} 14\hat{\beta}_0 + 60,72\hat{\beta}_1 + 3.155\hat{\beta}_2 = 17.495 \\ 60,72\hat{\beta}_0 + 264,26\hat{\beta}_1 + 13.683,50\hat{\beta}_2 = 75.738,30 \\ 3.155\hat{\beta}_0 + 13.683,50\hat{\beta}_1 + 716.425\hat{\beta}_2 = 4.001.120 \end{cases} \rightarrow \hat{\beta}_0 = -521,24, \quad \hat{\beta}_1 = -151,88 \quad \text{e} \quad \hat{\beta}_2 = 10,78.$$

Portanto:

$$\hat{y} = -521,24 - 151,88x_1 + 10,78x_2$$

Propriedade dos estimadores

Resumo



$$E(\hat{\beta}) = \beta$$



$$\text{Cov}(\hat{\beta}) = \sigma^2(X'X)^{-1}$$



$$SQE = Y'(I - X(X'X)^{-1}X')Y$$



$$\hat{\sigma}^2 = QME = \frac{SQE}{n - p - 1}$$

Propriedade dos estimadores

Teorema

Se $X \sim N_p(\mu; \Sigma)$, então, $X'AX \sim \chi^2_{r(A), \delta}$ (Qui-quadrado não central) se, e somente se, $A\Sigma$ é idempotente.

Em que:

- $r(A)$: representa o posto (rank) da matriz A, ou seja, o número de colunas linearmente independentes da matriz A.
- $\delta = \frac{1}{2}\mu'A\mu$ representa o parâmetro de não centralidade.
- Idempotente: $A\Sigma \times A\Sigma = A\Sigma$.

Assim, $\frac{SQE}{\sigma^2} = \frac{Y'}{\sigma} [I - X(X'X)^{-1}X'] \frac{Y}{\sigma} \sim \chi^2_{r[I - X(X'X)^{-1}X']; \delta}$

Pois $(I - X(X'X)^{-1}X')$ é idempotente. Além disso, temos que

$$\delta = \frac{1}{2} \frac{\beta'X'(I - X(X'X)^{-1}X')X\beta}{\sigma^2} = 0 \quad \text{e} \quad r(I - X(X'X)^{-1}X') = n - (p + 1).$$

Assumindo que $\varepsilon \sim N_p(0, \sigma^2 I_p)$, temos que $Y \sim N_p(X\beta, \sigma^2 I_p)$, então

$$\frac{Y}{\sigma} \sim N_p \left(\frac{X\beta}{\sigma}, I_p \right)$$

Exemplo

O ganho de um transistor consiste na diferença entre o emissor e o coletor. A variável **Ganho** (em hFE) pode ser controlada no processo de deposição de íons por meio das variáveis **Tempo de emissão** (em minutos) e **Dose de íons** ($\times 10^{14}$).

Observação	Tempo (min)	Dose de íons ($\times 10^{14}$)	Ganho (hFE)
1	195	4	1004
2	255	4	1636
3	195	4,6	852
4	255	4,6	1506
5	225	4,2	1272
6	225	4,1	1270
7	225	4,6	1269
8	195	4,3	903
9	255	4,3	1555
10	225	4	1260
11	225	4,7	1146
12	225	4,3	1276
13	225	4,72	1225
14	230	4,3	1321

$$(X'X) = \begin{bmatrix} 14 & 0,1667 & -0,8889 \\ 0,1667 & 6,0278 & -0,0278 \\ -0,8889 & -0,0278 & 7,0556 \end{bmatrix} \Rightarrow$$

$$(X'X)^{-1} = \begin{bmatrix} 0,0720 & -0,0019 & 0,0091 \\ -0,0019 & 0,1660 & 0,0004 \\ 0,0091 & 0,0004 & 0,1429 \end{bmatrix} \text{ e}$$

Portanto, as estimativas $\hat{\beta}$ são:

$$\hat{\beta} = (X'X)^{-1}X'y = \begin{bmatrix} 1242,31 \\ 323,43 \\ -54,77 \end{bmatrix}.$$

$$X'y = \begin{bmatrix} 17495 \\ 2158,167 \\ -1499,722 \end{bmatrix}.$$

Assim, a equação da regressão é dada por:

$$\hat{y} = 1242,31 + 323,43 x_1 - 54,77 x_2$$

ANOVA

Análise de Variância

Análise de Variância

Soma de Quadrados



$$SQT = Y'Y - \frac{Y'JY}{n}$$

$$SQR = Y'\left(H - \frac{J}{n}\right)Y$$

$$SQE = Y'Y - \hat{\beta}'X'Y$$

Sendo que $H = X(X'X)^{-1}X'$ é denominada matriz hat (chapéu) e

$$J = \begin{bmatrix} 1 & 1 & \dots & 1 \\ 1 & 1 & \dots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \dots & 1 \end{bmatrix}_{n \times n}.$$

ANOVA



Fonte	Soma de quadrados	GL	Quadrado médio
Rregressão	SQR	p	$\frac{SQR}{p}$
Erro (Resíduo)	SQE	$n - p - 1$	$\frac{SQE}{(n-p-1)}$
Total	SQT	$n - 1$	

Análise de Variância

Teste de hipótese

$$\begin{cases} H_0 : \beta_1 = \beta_2 = \dots = \beta_p = 0 \\ H_1 : \beta_j \neq 0 \text{ para qualquer } j = 1, \dots, p \end{cases}$$

Estatística do Teste

$$F_0 = \frac{\frac{SQR}{p}}{\frac{SQE}{n-p-1}} = \frac{QMR}{QME} \sim F_{(p,n-p-1)}$$

O valor de F_0 deve ser comparado com o quantil distribuição **F de Snedecor** com p graus de liberdade no numerador e $n-p-1$ graus de liberdade no denominador para um determinado nível de significância α

- Se $F_0 \leq F_{(p;n-p-1;1-\alpha)}$ não rejeitamos H_0 , ou seja, rejeitamos a hipótese de que o coeficiente angular seja significativo;
- Se $F_0 > F_{(p;n-p-1;1-\alpha)}$ rejeitamos H_0 , ou seja, não rejeitamos a hipótese de que o coeficiente angular seja significativo;

- Outra forma de definirmos um critério para avaliarmos o teste de hipótese é o P-valor.
- O P-valor = $P(F_{(p;n-p-1)} > F_0)$ representa o menor nível de significância para o qual rejeitamos H_0 . Logo, para um nível de significância $\alpha = 0,05$ adotado (por exemplo), rejeitamos H_0 se o P-valor obtido for menor que 0,05, enquanto que não rejeitamos H_0 se o P-valor for maior que 0,05.

Fonte	Soma de quadrados	GL	Média dos quadrados	F_0
Regressão	SQR	p	$\frac{SQR}{p}$	$\frac{QMR}{QME}$
Erro (Resíduo)	SQE	$n - p - 1$	$\frac{QME}{(n-p-1)}$	
Total	SQT	$n - 1$		

Exemplo

O ganho de um transistor consiste na diferença entre o emissor e o coletor. A variável **Ganho** (em hFE) pode ser controlada no processo de deposição de íons por meio das variáveis **Tempo de emissão** (em minutos) e **Dose de íons** ($\times 10^{14}$).

Observação	Tempo (min)	Dose de íons ($\times 10^{14}$)	Ganho (hFE)
1	195	4	1004
2	255	4	1636
3	195	4,6	852
4	255	4,6	1506
5	225	4,2	1272
6	225	4,1	1270
7	225	4,6	1269
8	195	4,3	903
9	255	4,3	1555
10	225	4	1260
11	225	4,7	1146
12	225	4,3	1276
13	225	4,72	1225
14	230	4,3	1321

Soma de Quadrados:

$$\begin{aligned}
 SQT &= y'y - \frac{\left(\sum_{i=1}^{14} y_i\right)^2}{14} \\
 &= 22567889 - \frac{(17495)^2}{14} \\
 &= 665387,2 \\
 SQR &= \hat{\beta}'X'y - \frac{\left(\sum_{i=1}^{14} y_i\right)^2}{14} \\
 &= 22514467,9 - 21862501,8 \\
 &= 651966,1 \\
 SQE &= S_{yy} - SQR \\
 &= 665387,2 - 651966,1 \\
 &= 13421,1
 \end{aligned}$$

Assim, a equação da regressão é dada por:

Fonte	Soma de quadrados	GL	Quadrado médio	F ₀	P-valor
Regressão	651966,1	2	$\frac{651966,1}{2} = 325983,0$	267,18	0,00
Erro (Resíduo)	13421,1	11	$\frac{13421,1}{11} = 1220,1$		
Total	665387,2	13			

Para $\alpha = 0,05$, temos $F_0 = 267,18 > F_{(0,95;2;11)} = 3,98$ e P-valor = $P[F_{2;11} > F_0] = 0$. Assim rejeitamos H_0 .

Medidas de Associação

Coeficiente de determinação R^2



$$R^2 = \frac{SQR}{SQT} = 1 - \frac{SQE}{SQT} \text{ com } 0 \leq R^2 \leq 1$$

$$R_{\text{ajustado}}^2 = 1 - \left(\frac{n-1}{n-p-1} (1 - R^2) \right)$$

É uma medida descritiva da qualidade do ajuste obtido, em geral, referimos como a quantidade de variabilidade nos dados que é explicada pelo modelo de regressão ajustado

o valor do coeficiente de determinação depende do número de observações (n), tendendo a crescer quando n diminui. Se $n = 2$ temos sempre $R^2 = 1$

é sempre possível torná-lo maior pela adição de um número suficiente de termos ao modelo. Por exemplo, se não há dados repetidos (mais do que um valor para um mesmo x) um polinômio de grau ($n - 1$) dará um ajuste perfeito ($R^2 = 1$) para n dados.

Embora (R^2) aumente com a adição de termos ao modelo, isto não significa necessariamente que o novo modelo é superior ao anterior. A menos que a soma de quadrados residual do novo modelo seja reduzida por uma quantidade igual ao quadrado médio residual original, o novo modelo terá um quadrado médio residual maior do que o original, devido a perda de 1 grau de liberdade.

Precauções no uso R^2

Exemplo

O ganho de um transistor consiste na diferença entre o emissor e o coletor. A variável **Ganho** (em hFE) pode ser controlada no processo de deposição de íons por meio das variáveis **Tempo de emissão** (em minutos) e **Dose de íons** ($\times 10^{14}$).

Observação	Tempo (min)	Dose de íons ($\times 10^{14}$)	Ganho (hFE)
1	195	4	1004
2	255	4	1636
3	195	4,6	852
4	255	4,6	1506
5	225	4,2	1272
6	225	4,1	1270
7	225	4,6	1269
8	195	4,3	903
9	255	4,3	1555
10	225	4	1260
11	225	4,7	1146
12	225	4,3	1276
13	225	4,72	1225
14	230	4,3	1321

$$R^2 = \frac{SQR}{SQT} = \frac{651966,1}{665387,2} = 0,9798.$$

$$\begin{aligned} R^2_{ajustado} &= 1 - \left(\frac{n-1}{n-p-1} (1-R^2) \right) \\ &= 1 - \left(\frac{13}{11} \right) (1-0,9798) \\ &= 0,9762. \end{aligned}$$

Fonte	Soma de quadrados	GL	Quadrado médio	F ₀	P-valor
Regressão	651966,1	2	$\frac{651966,1}{2} = 325983,0$	267,18	0,00
Erro (Resíduo)	13421,1	11	$\frac{13421,1}{11} = 1220,1$		
Total	665387,2	13			

Testes individuais para os coeficientes da regressão

Inferência para os coeficientes da regressão

As hipóteses para testar a significância de qualquer coeficiente de regressão é dado por:

Estatística do Teste

$$\begin{cases} H_0: \beta_j = 0 \\ H_1: \beta_j \neq 0 \end{cases} \quad j = 0, 1, \dots, p$$

$$T_0 = \frac{\hat{\beta}_j - \beta_j}{\sqrt{\hat{\sigma}^2 C_{jj}}} \sim t_{(n-p-1)} \text{ (Sob } H_0)$$

O valor de T_0 deve ser comparado com o quantil distribuição t- Student com $n-p-1$ graus de liberdade para um determinado nível de significância α

- Se $|T_0| \leq t_{(n-p-1; 1-\frac{\alpha}{2})}$ não rejeitamos H_0 , ou seja, rejeitamos a hipótese de que o coeficiente angular seja significativo;
- Se $|T_0| > t_{(n-p-1; 1-\frac{\alpha}{2})}$ rejeitamos H_0 , ou seja, não rejeitamos a hipótese de que o coeficiente angular seja significativo;

- Outra forma de definirmos um critério para avaliarmos o teste de hipótese é o P-valor.
- O P-valor = $2 * P(t_{(n-p-1)} > |T_0|)$ representa o menor nível de significância para o qual rejeitamos H_0 . Logo, para um nível de significância $\alpha = 0,05$ adotado (por exemplo), rejeitamos H_0 se o P-valor obtido for menor que 0,05, enquanto que não rejeitamos H_0 se o P-valor for maior que 0,05.

Intervalo de Confiança para β_j com $(1 - \alpha)100\%$ é dado por:

$$\left[\hat{\beta}_j - t_{(1-\frac{\alpha}{2}; n-p-1)} \sqrt{\hat{\sigma}^2 C_{jj}} ; \hat{\beta}_j + t_{(1-\frac{\alpha}{2}; n-p-1)} \sqrt{\hat{\sigma}^2 C_{jj}} \right]$$

Exemplo

O ganho de um transistor consiste na diferença entre o emissor e o coletor. A variável **Ganho** (em hFE) pode ser controlada no processo de deposição de íons por meio das variáveis **Tempo de emissão** (em minutos) e **Dose de íons** ($\times 10^{14}$).

$$(X'X) = \begin{bmatrix} 14 & 0,1667 & -0,8889 \\ 0,1667 & 6,0278 & -0,0278 \\ -0,8889 & -0,0278 & 7,0556 \end{bmatrix} \Rightarrow C = (X'X)^{-1} = \begin{bmatrix} 0,0720 & -0,0019 & 0,0091 \\ -0,0019 & 0,1660 & 0,0004 \\ 0,0091 & 0,0004 & 0,1429 \end{bmatrix}.$$

Para
 $H_0: \beta_0 = 0$

$$t_0 = \frac{\hat{\beta}_0}{\sqrt{\hat{\sigma}^2 C_{00}}} = \frac{1242,31}{\sqrt{(1220,1)(0,0720)}} = \frac{1242,31}{9,37} = 132,58$$

Para
 $H_0: \beta_1 = 0$

$$t_0 = \frac{\hat{\beta}_1}{\sqrt{\hat{\sigma}^2 C_{11}}} = \frac{323,43}{\sqrt{(1220,1)(0,1660)}} = \frac{323,43}{14,2} = 22,77$$

Para
 $H_0: \beta_2 = 0$

$$t_0 = \frac{\hat{\beta}_2}{\sqrt{\hat{\sigma}^2 C_{22}}} = \frac{-54,77}{\sqrt{(1220,1)(0,1429)}} = \frac{-54,77}{13,2} = -4,15$$

P-valor

$$2 * P(t_{11} > |132,58|) = 0$$

$$2 * P(t_{11} > |22,77|) = 0$$

$$2 * P(t_{11} > |-4,15|) = 0,00162$$

Como todos valores absolutos destas estatísticas são maiores que o valor crítico $t_{(0,025;11)} = 2,201$ as hipóteses H_0 são rejeitadas. Notamos que os p-valores são menores que $\alpha = 0,05$.

Intervalo de Confiança para β_1 com $(1 - \alpha)100\%$ é dado por:

$$\left[\hat{\beta}_1 - t_{(1-\frac{\alpha}{2};n-p-1)} \sqrt{\hat{\sigma}^2 C_{jj}} ; \hat{\beta}_1 + t_{(1-\frac{\alpha}{2};n-p-1)} \sqrt{\hat{\sigma}^2 C_{jj}} \right] = [292,1 ; 354,75]$$

Intervalo de Confiança para resposta média

Resposta média (denotada por $E(y | x_0)$) em um ponto particular da amostra, $x'_0 = (1, x_{01}, x_{02}, \dots, x_{0p})$.

A resposta média para este ponto é,

$$\mu_{y|x_0} = \beta_0 + \beta_1 x_{01} \beta_2 x_{02} + \cdots + \beta_p x_{0p} = x'_0 \beta \xrightarrow{\text{estimativa}} \hat{\mu}_{y|x_0} = x'_0 \hat{\beta}$$

A variância de $\hat{\mu}_{y|x_0}$ é dada por

$$Var(\hat{\mu}_{y|x_0}) = \sigma^2 x'_0 (X'X)^{-1} x_0$$

Assim, temos

$$\frac{\hat{\mu}_{y|x_0} - \mu_{y|x_0}}{\sqrt{Var(\hat{\mu}_{y|x_0})}} \sim t_{(n-p-1)}$$

Intervalo de Confiança para a resposta média no ponto x_0 com $(1 - \alpha)100\%$ é dado por:

$$\left[\hat{\mu}_{y|x_0} - t_{(\frac{\alpha}{2}; n-p-1)} \sqrt{\sigma^2 x'_0 (X'X)^{-1} x_0} ; \hat{\mu}_{y|x_0} + t_{(\frac{\alpha}{2}; n-p-1)} \sqrt{\sigma^2 x'_0 (X'X)^{-1} x_0} \right]$$

Exemplo

O ganho de um transistor consiste na diferença entre o emissor e o coletor. A variável **Ganho** (em hFE) pode ser controlada no processo de deposição de íons por meio das variáveis **Tempo de emissão** (em minutos) e **Dose de íons** ($\times 10^{14}$).

Em termos das variáveis codificadas x_{01} e x_{02} , corresponde a $x_{01} = 0$ e $x_{02} = 0$, então $x'_0 = [1 \ 0 \ 0]$.

A estimativa da resposta média neste ponto é:

$$\hat{\mu}_{y|x_0} = x'_0 \hat{\beta} = [1 \ 0 \ 0] \begin{bmatrix} 1243,31 \\ 323,43 \\ -54,77 \end{bmatrix} = 1242,31$$

A variância de $\hat{\mu}_{y|x_0}$ é dada por

$$\begin{aligned} Var(\hat{\mu}_{y|x_0}) &= \sigma^2 x'_0 (X'X)^{-1} x_0 \\ &= \sigma^2 [1 \ 0 \ 0] \begin{bmatrix} 0,0720 & -0,0019 & 0,0091 \\ -0,0019 & 0,1660 & 0,0004 \\ 0,0091 & 0,0004 & 0,1429 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} \end{aligned}$$

Sabemos que o estimador de σ^2 é $\hat{\sigma}^2 = QME = 1220,1$.

Intervalo de Confiança para a resposta média no ponto x_0 com $(1 - \alpha)100\%$ é dado por:

$$\begin{aligned} &\left[\hat{\mu}_{y|x_0} - t_{\left(\frac{\alpha}{2}; n-p-1\right)} \sqrt{\sigma^2 x'_0 (X'X)^{-1} x_0} ; \hat{\mu}_{y|x_0} + t_{\left(\frac{\alpha}{2}; n-p-1\right)} \sqrt{\sigma^2 x'_0 (X'X)^{-1} x_0} \right] \\ &= [1221,8 ; 1262,94] \end{aligned}$$

Previsões para novas observações

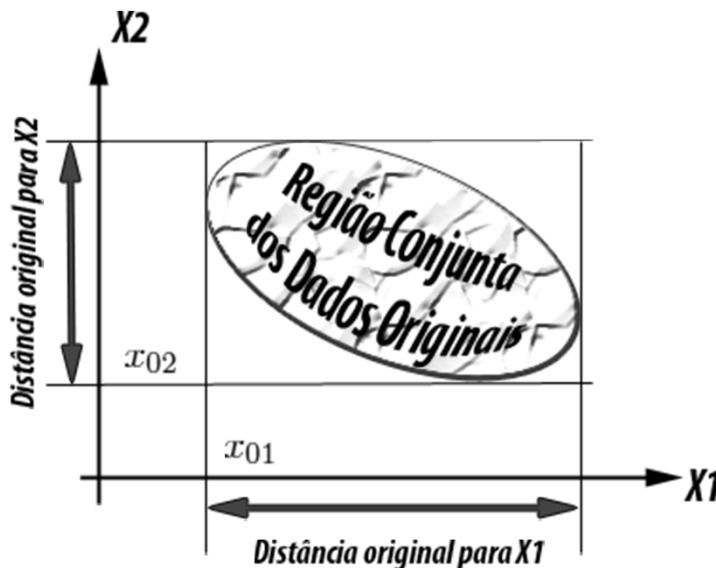
Previsões para novas observações

Se, $x'_0 = [1 \ x_{01} \ x_{02} \ \dots \ x_{0p}]$, o valor predito para a nova observação é dada por:

$$\hat{y}(x_0) = x'_0 \hat{\beta}$$

Intervalo de previsão com $(1 - \alpha)100\%$ para futuras observações é dado por:

$$\left[\hat{y}(x_0) - t_{(\frac{\alpha}{2}; n-p-1)} \sqrt{\sigma^2(1 + x'_0(X'X)^{-1}x_0)} ; \hat{y}(x_0) + t_{(\frac{\alpha}{2}; n-p-1)} \sqrt{\sigma^2(1 + x'_0(X'X)^{-1}x_0)} \right]$$



Observações

- É preciso cuidado para estimar a variável resposta para pontos que estão além da região da amostra.
- É bem provável que um modelo ajustado dentro da região dos dados possa não estar mais ajustado quando usarmos observações fora dessa região.
- Em regressão múltipla é fácil extrapolar novos dados mas para tal devemos ter certeza que os níveis das variáveis definam uma região conjunta que contenha os dados.

Exemplo

O ganho de um transistor consiste na diferença entre o emissor e o coletor. A variável **Ganho** (em hFE) pode ser controlada no processo de deposição de íons por meio das variáveis **Tempo de emissão** (em minutos) e **Dose de íons** ($\times 10^{14}$).

Vamos encontrar um intervalo de predição com 95% de confiança para as próximas observações, do ganho do transistor, no ponto de $x_{01} = 225$ min e $x_{02} = 4,36 \times 10^{14}$ íons.

$$\hat{y}(x_0) = x_0' \hat{\beta} = [1, 0, 0] \begin{bmatrix} 1243,31 \\ 323,43 \\ -54,77 \end{bmatrix} = 1242,31.$$

Do exemplo anterior obtemos que:

$$x_0'(X'X)^{-1}x_0 = 0,072,$$

Intervalo de previsão com $(1 - \alpha)100\%$ para futuras observações é dado por:

$$\left[\hat{y}(x_0) - t_{(\frac{\alpha}{2}; n-p-1)} \sqrt{\sigma^2(1 + x_0'(X'X)^{-1}x_0)} ; \hat{y}(x_0) + t_{(\frac{\alpha}{2}; n-p-1)} \sqrt{\sigma^2(1 + x_0'(X'X)^{-1}x_0)} \right] \\ = [1162,71 ; 1321,91]$$

Análise de Resíduos

Análise dos Resíduos

O resíduo (e_i) é dado pela diferença entre a variável resposta (Y_i) observada e a variável resposta estimada (\hat{Y}_i)

$$e_i = Y_i - \hat{Y}_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{1i} - \cdots - \hat{\beta}_p x_{pi}, \\ i = 1, \dots, n$$

Suposições para $Y = X\beta + \varepsilon$:

- O modelo é linear
- O erro tem média zero
- O erro tem variância σ^2 , constante
- Os erros são não correlacionados, ou seja,
 $Cov(\varepsilon_i, \varepsilon_j) = 0, (i \neq j)$
- Os erros $\varepsilon_i \sim N(0, \sigma^2)$

Algumas técnicas gráficas para análise dos resíduos



- Gráfico dos resíduos versus valores ajustados: verifica a homoscedasticidade do modelo, isto é, σ^2 constante.
- Gráfico dos resíduos versus a ordem de coleta dos dados: avaliar a hipótese de independência dos dados.
- Papel de probabilidade normal: verificar a normalidade dos dados.
- Gráfico dos Resíduos Studentizados versus valores ajustados: verifica se existem outliers em Y.
- Gráfico dos Resíduos Padronizados versus valores ajustados: verifica se existem outliers em Y.
- Gráfico do Leverage (Diagonal da Matriz H): verifica se existem outliers em X.

Para a análise formal dos resíduos, podemos realizar os seguintes testes



- Teste de Normalidade em que detalhes estão contidos no conteúdo de Inferência.
- Teste de Durbin-Watson para testar independência dos resíduos.
- Teste de Breusch-Pagan e Goldfeld-Quandt para testar se os resíduos são homoscedásticos.
- Teste de falta de ajuste para verificar se o modelo ajustado é realmente linear.

Diagnóstico de Outliers

Outlier é uma observação extrema, ou seja, é um ponto com comportamento diferente dos demais. Além de diagnosticar heteroscedasticidade, o gráfico de resíduos versus valores ajustados também auxilia na detecção de pontos atípicos.

A detecção de pontos atípicos tem por finalidade identificar:

- Outliers com relação a X
- Outliers com relação a Y
- Observações influentes

A seguir apresentamos algumas técnicas utilizadas:

Outliers em Y

- Resíduos Padronizados
- Resíduos Studentizados

Outliers em X

- Matriz chapéu H
- h_{ii} é o i-ésimo elemento da diagonal principal da matriz H e também é chamado de leverage da observação i.

Pontos Influentes

- DFFITS
- DFBETA
- Distância de Cook

Outliers em Y

Resíduos Padronizados

O resíduo padronizado, d_i , corresponde ao resíduo bruto dividido pelo erro padrão estimado dos resíduos, \sqrt{QME}

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

Se os erros têm distribuição normal, então aproximadamente 95% dos resíduos padronizados (d_i), devem estar no intervalo de $(-2,2)$. Resíduos fora desse intervalo podem indicar a presença de outliers.

Resíduos Studentizados

Vetor de resíduos

$$e = Y - \hat{Y} = Y - X\hat{\beta} = Y - HY = (I - H)Y$$

A matriz de covariâncias dos resíduos é,

$$Cov[e] = \sigma^2(I - H)$$

Resíduos Studentizados

$$r_i = \frac{e_i}{\sqrt{QME(I - h_{ii})}}, \quad i = 1, 2, \dots, n \text{ e } \hat{\sigma}^2 = QME$$

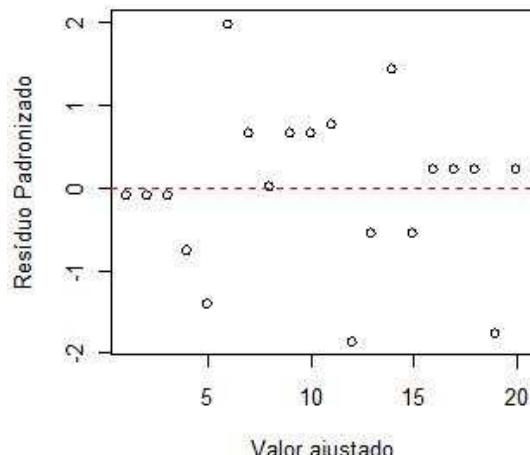
Exemplo (Motivação 1)

Em problemas de tratamento térmico deseja-se estabelecer uma relação entre a temperatura da estufa e uma característica da qualidade (Dureza). Determinar os valores de temperatura em C que “otimizam” a performance do processo de tratamento térmico em relação a estrutura metalográfica do material, avaliada em relação de dureza em HB.

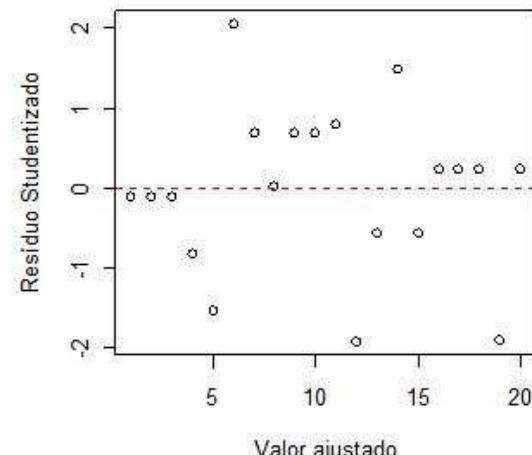
Outlier em Y

e	Studentizados	Padronizados
-0,14	-0,100	-0,09
-0,14	-0,100	-0,093
-0,14	-0,100	-0,093
-1,14	-0,813	-0,754
-2,14	-1,526	-1,415
3,02	2	2
1,02	1	1
0,02	0,014	0,013
1,02	0,696	0,675
1,02	0,696	0,675
1,18	0,805	0,780
-2,82	-1,923	-1,865
-0,82	-1	-1
2,18	1,487	1,442
-0,82	-1	-1
0,34	0,242	0,225
0,34	0,242	0,225
0,34	0,242	0,225
-2,66	-1,897	-1,759
0,34	0,244	0,22

Res. Padronizado vs Valores Ajustados



Res. Studentizado vs Valores Ajustados



Teste de Outliers

Observações	T-Valor	P-valorBonferroni
6	2,289843184	0,701817143

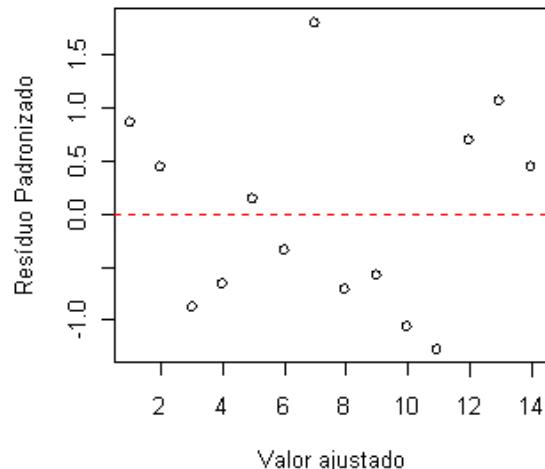
Exemplo (Motivação 2)

O ganho de um transistor consiste na diferença entre o emissor e o coletor. A variável **Ganho** (em hFE) pode ser controlada no processo de deposição de íons por meio das variáveis **Tempo de emissão** (em minutos) e **Dose de íons** ($\times 10^{14}$).

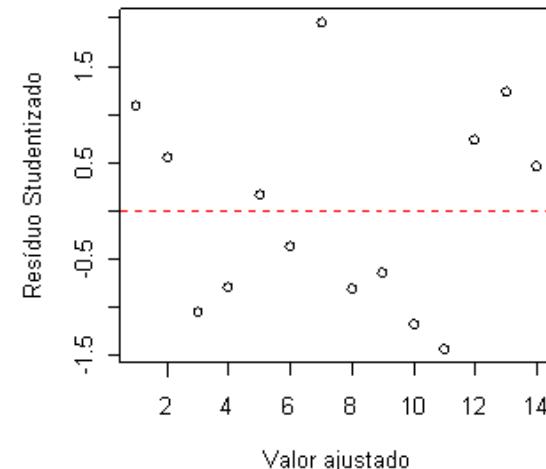
Outlier em Y

e	Studentizados	Padronizados
30,35	1,092	0,869
15,48	0,553	0,443
-30,36	-1,052	-0,869
-23,23	-0,801	-0,665
5,34	0,16	0,153
-11,87	-0,365	-0,34
63,2	1,96	1,809
-25,01	-0,823	-0,716
-19,88	-0,651	-0,569
-37,09	-1,185	-1,062
-44,58	-1,442	-1,276
24,56	0,73	0,703
37,46	1,225	1,072
15,65	0,466	0,448

Res. Padronizado vs Valores Ajustados



Res. Studentizado vs Valores Ajustados



Teste de Outliers

Observações	T-Valor	-valorBonferroni
7	2,31619333	0,602713713

Outliers em X

Outliers em X

- Matriz chapéu H
- h_{ii} é o i-ésimo elemento da diagonal principal da matriz H e também é chamado de leverage da observação i.

Leverage

$$h_{ii} = x_i'(X'X)^{-1}x_i$$

Em que x_i' é a linha da Matriz X, correspondente a i-ésima observação.

Observações com alta leverage
(outliers em X)

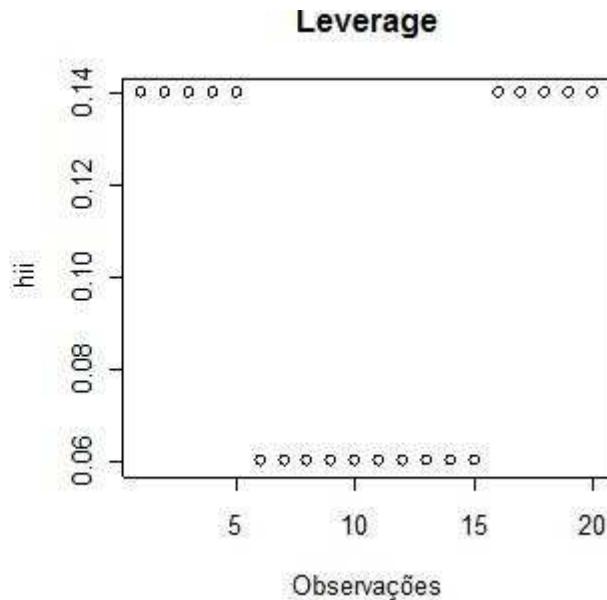
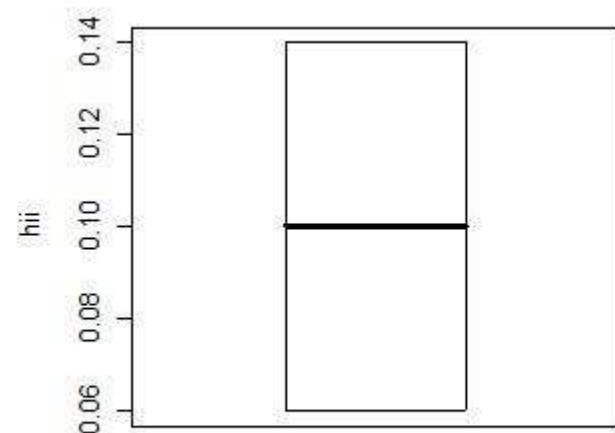
- Observar se há valores extremos de h_{ii} em um boxplot.
- Se a amostra não é grande, considerar h_{ii} grande se é maior que duas vezes a média dos h_{ii} . Como $\sum_{i=1}^n h_{ii} = \text{posto}(X) = p + 1$, a média dos h_{ii} é $p+1/n$. É recomendado destacar as observações para as quais $h_{ii} > 2\frac{p+1}{n}$.
- Indicar leverage muito alta se $h_{ii} > 0,5$, se a amostra for considerada grande.

Exemplo (Motivação 1)

Em problemas de tratamento térmico deseja-se estabelecer uma relação entre a temperatura da estufa e uma característica da qualidade (Dureza). Determinar os valores de temperatura em C que “otimizam” a performance do processo de tratamento térmico em relação a estrutura metalográfica do material, avaliada em relação de dureza em HB.

Outlier em X

h_{ii}	Valores
h_1	0,14
h_2	0,14
h_3	0,14
h_4	0,14
h_5	0,14
h_6	0,06
h_7	0,06
h_8	0,06
h_9	0,06
h_{10}	0,06
h_{11}	0,06
h_{12}	0,06
h_{13}	0,06
h_{14}	0,06
h_{15}	0,06
h_{16}	0,14
h_{17}	0,14
h_{18}	0,14
h_{19}	0,14
h_{20}	0,14

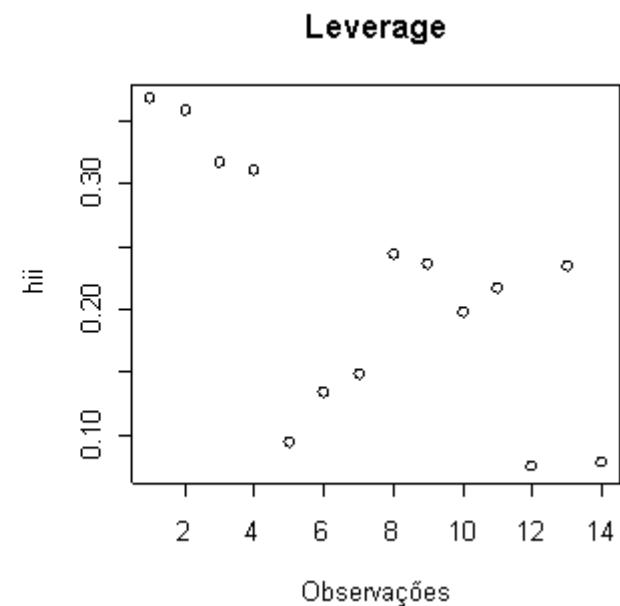
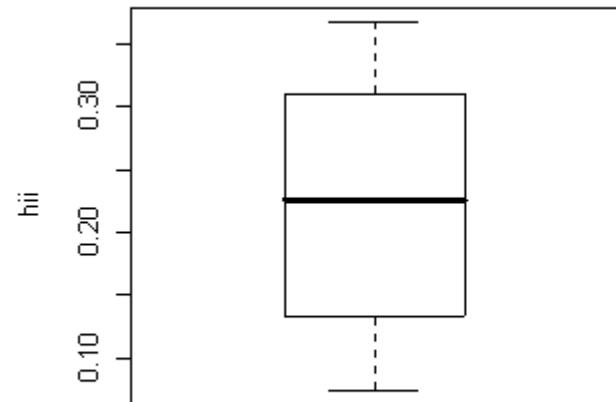


Exemplo (Motivação 2)

O ganho de um transistor consiste na diferença entre o emissor e o coletor. A variável **Ganho** (em hFE) pode ser controlada no processo de deposição de íons por meio das variáveis **Tempo de emissão** (em minutos) e **Dose de íons** ($\times 10^{14}$).

Outlier em X

h_{ii}	Valores
h_1	0,36744
h_2	0,35801
h_3	0,3169
h_4	0,3102
h_5	0,09219
h_6	0,1334
h_7	0,1476
h_8	0,2429
h_9	0,2348
h_{10}	0,19677
h_{11}	0,216595
h_{12}	0,072974
h_{13}	0,233037
h_{14}	0,07691



Pontos Influentes

DFFITS



- mede a influência que a observação i tem sobre seu próprio valor ajustado.
 - Consideremos a medida $DFFITS_{(i)} = \frac{\hat{Y}_i - \hat{Y}_{i(i)}}{\sqrt{QME_{(i)} h_{ii}}}$
 - essa técnica mede o quanto a inclusão da observação i aumenta ou diminui seu valor predito.
- Dizemos que um ponto outlier é influente segundo o DFFITS se
- Para amostras pequenas ou médias $|DFFITS_{(i)}| > 1$
 - Para amostras grandes $|DFFITS_{(i)}| > 2\sqrt{\frac{p}{n}}$

DFBETA



- mede a influência da observação i sobre o coeficiente de X_j .
 - Consideremos a medida $DFBETA_{j(i)} = \frac{\hat{\beta}_j - \hat{\beta}_{j(i)}}{\sqrt{QME_i c_{ii}}}, \quad j = 0, 1, \dots, p$
- São consideradas observações influentes aquelas que
- Para amostras pequenas $|DFBETA| > 1$
 - Para amostras grandes $|DFBETA| > \frac{2}{\sqrt{n}}$

DFBETA

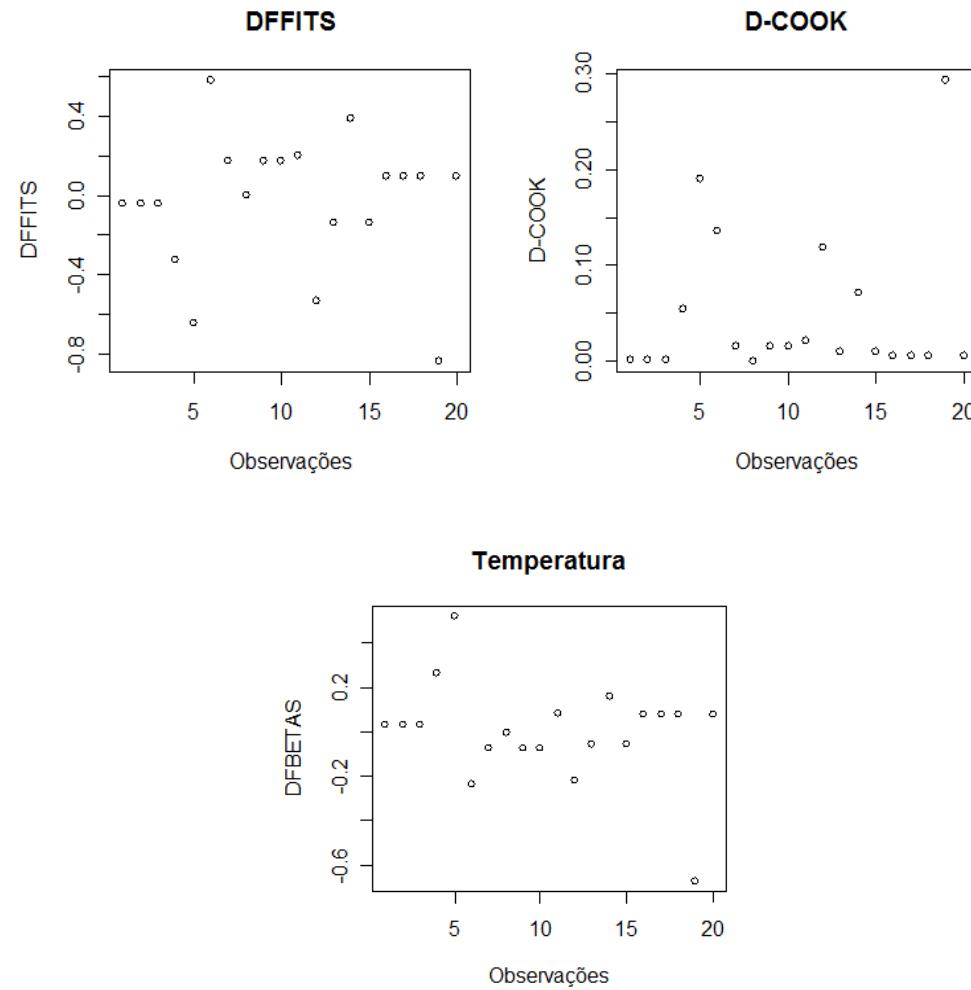


- A distância de Cook mede a influência da observação i sobre todos os valores ajustados \hat{Y}_i .
- Consideremos a medida $D_i = \frac{e_i^2}{p QME} \frac{h_{ii}}{(1-h_{ii})^2}$
- Percebemos que D_i é grande quando ou o resíduo e_i é grande, a leverage h_{ii} é grande ou ambos. Destacamos as observações quando $D_i > 1$.

Exemplo (Motivação 1)

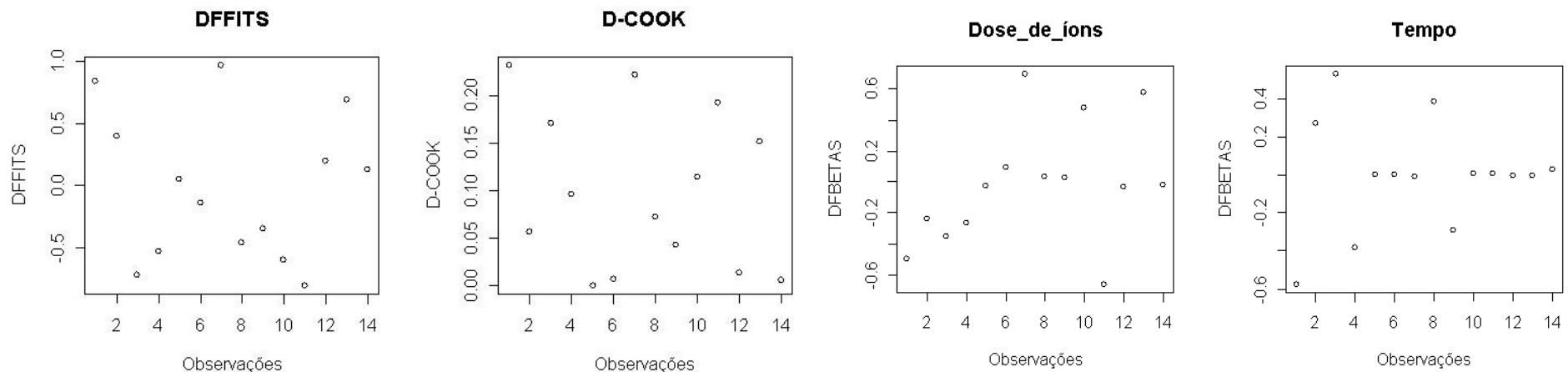
Em problemas de tratamento térmico deseja-se estabelecer uma relação entre a temperatura da estufa e uma característica da qualidade (Dureza). Determinar os valores de temperatura em C que “otimizam” a performance do processo de tratamento térmico em relação a estrutura metalográfica do material, avaliada em relação de dureza em HB.

Observações	DFFIT	$DF\beta_0$	$DF\beta_1$	DCOOK
1	-0,039	-0,032	0,031	8,11E-04
2	-0,039	-0,032	0,031	8,11E-04
3	-0,039	-0,032	0,031	8,11E-04
4	-0,325	-0,265	0,260	5,38E-02
5	-0,641	-0,523	0,514	1,90E-01
6	0,579	0,249	-0,236	1,35E-01
7	0,173	0,075	-0,071	1,54E-02
8	0,003	0,001	-0,001	5,94E-06
9	0,173	0,075	-0,071	1,54E-02
10	0,173	0,075	-0,071	1,54E-02
11	0,201	-0,078	0,082	2,07E-02
12	-0,530	0,204	-0,216	1,18E-01
13	-0,139	0,053	-0,057	9,98E-03
14	0,390	-0,150	0,159	7,06E-02
15	-0,139	0,053	-0,057	9,98E-03
16	0,095	-0,075	0,076	4,78E-03
17	0,095	-0,075	0,076	4,78E-03
18	0,095	-0,075	0,076	4,78E-03
19	-0,831	0,654	-0,667	2,93E-01
20	0,095	-0,075	0,076	4,78E-03



Exemplo (Motivação 2)

O ganho de um transistor consiste na diferença entre o emissor e o coletor. A variável **Ganho** (em hFE) pode ser controlada no processo de deposição de íons por meio das variáveis **Tempo de emissão** (em minutos) e **Dose de íons** ($\times 10^{14}$).



Observação	DFFITS	DF β_0	DF β_1	DF β_2	DCOOK
1	0,841	0,745	-0,492	-0,573	0,231
2	0,399	0,055	-0,236	0,268	0,057
3	-0,720	-0,018	-0,352	0,526	0,171
4	-0,528	0,418	-0,262	-0,382	0,096
5	0,049	0,022	-0,023	-0,001	0,001
6	-0,137	-0,084	0,094	0,002	0,007
7	0,964	-0,535	0,692	-0,010	0,222
8	-0,459	-0,257	0,037	0,384	0,072
9	-0,351	0,130	0,027	-0,291	0,043
10	-0,599	-0,417	0,478	0,008	0,115
11	-0,803	0,518	-0,657	0,007	0,192
12	0,200	0,036	-0,029	-0,004	0,014
13	0,693	-0,456	0,577	-0,005	0,152
14	0,130	0,005	-0,018	0,029	0,006

Diagnóstico de Homoscedasticidade

Homoscedasticidade é o termo para designar variância constante dos erros ε_i para observações diferentes.

Caso a suposição de homoscedasticidade não seja válida, podemos listar alguns efeitos no ajuste do modelo:

- Os erros padrões dos estimadores, obtidos pelo Método dos Mínimos Quadrados, são incorretos e portanto a inferência estatística não é válida.
- Não podemos mais dizer que os Estimadores de Mínimos Quadrados são os melhores estimadores de mínima variância para , embora ainda possam ser não viciados.

A seguir apresentamos algumas técnicas utilizadas:

Teste de Goldfeld-Quandt

- é utilizado para testar a homoscedasticidade dos resíduos. Entre as limitações deste teste está a exigência de que a amostra seja relativamente grande.
- O teste consiste inicialmente em ordenar as observações de acordo com a variável explicativa que se acredita a responsável pela heteroscedasticidade.
- Após isso, divide-se a amostra ordenada em 3 partes de tal forma que a parte do meio tenha aproximadamente 20% dos dados e que as partes 1 e 3 tenham quantidade de dados semelhantes.
- Então, ajusta-se um modelo de regressão com os dados da parte 1 e outro modelo de regressão com os dados da parte 3
- testa-se a hipótese de que as variâncias dos erros em ambas regressões são iguais contra a hipótese de que a variância dos erros na parte 3 é maior do que a variância dos erros na parte 1, utilizando o teste F.

A estatística de teste neste caso é dada por $F_{GQ} = \frac{SQE^b / (n_3 - (p+1))}{SQE^a / (n_1 - (p+1))}$

Técnica gráfica

Resíduos vs. Valores Ajustados

Diagnóstico de Independência

Para verificar se os resíduos são independentes, podemos utilizar técnicas gráficas e o testes de Durbin-Watson.

A seguir apresentamos algumas técnicas utilizadas:

Teste de Durbin-Watson

- é utilizado para detectar a presença de auto correlação (dependência) nos resíduos de uma análise de regressão.
- teste é baseado na suposição de que os erros no modelo de regressão são gerados por um processo autoregressivo de primeira ordem, de acordo com

$$\varepsilon_i = \rho \varepsilon_{i-1} + a_i$$

em que é o termo do erro do modelo na i-ésima observação, $a_i \stackrel{iid}{\sim} N(0, \sigma_a^2)$ e ρ ($|\rho| < 1$) é o parâmetro de autocorrelação.

- Testamos a presença de autocorrelação por meio das hipóteses

$$\begin{cases} H_0: \rho = 0 \\ H_1: \rho \neq 0 \end{cases}$$

- Sendo o resíduo associado à i-ésima observação, temos que a estatística do teste de Durbin-Watson é dada por

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

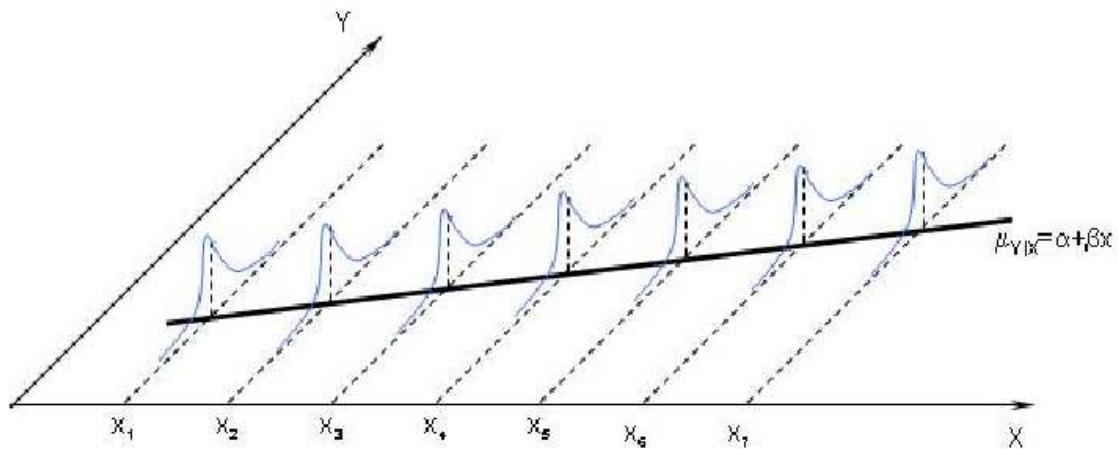
em que $0 \leq dw \leq 4$

Técnica gráfica

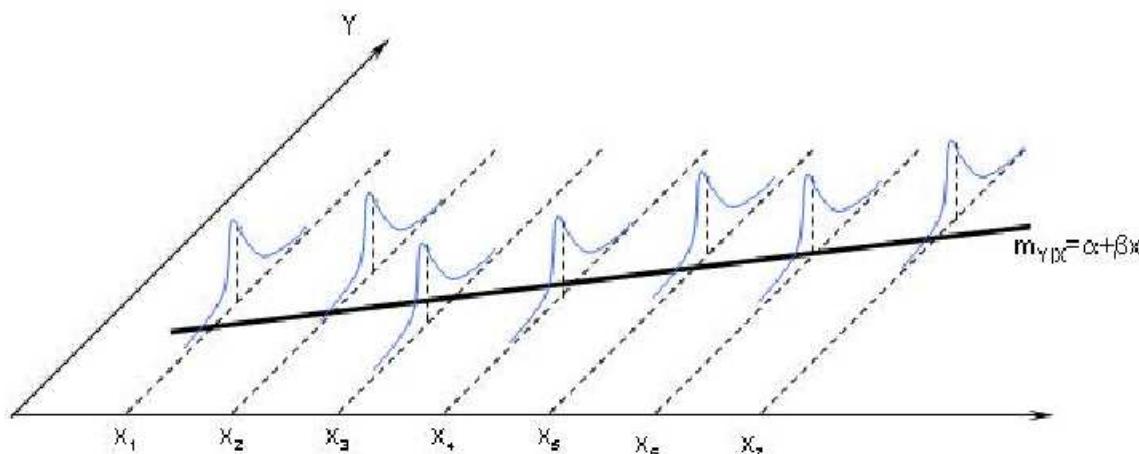
Resíduos vs. Ordem de Coleta

Teste da Falta de Ajuste (Lack of Fit)

Após o ajuste, é importante verificar se o modelo linear é adequado.



Reta de regressão
perfeitamente ajustada
sem Falta de Ajuste



Reta de regressão com
Falta de Ajuste

Teste da Falta de Ajuste (Lack of Fit)

A seguir, vamos fazer uma análise da falta de ajuste nos modelos simples e múltiplo. No caso simples, nosso modelo é:

$$y_{ij} = \beta_0 + \beta_1 x_i + \varepsilon \quad \begin{cases} i = 1, 2, \dots, m \\ j = 1, 2, \dots, n_i \end{cases}$$

Em que
□ n_i : representa o número de observações para o i -ésimo valor de x .

Supondo que temos m diferentes valores da variável explicativa (x_1, x_2, \dots, x_m) e que temos n_i réplicas da variável resposta para cada valor da variável explicativa, ou seja,

$$\left\{ \begin{array}{rcl} x_1 & \Rightarrow & y_{11} \quad y_{12} \quad y_{13} \quad \dots \quad y_{1n_1} \\ x_2 & \Rightarrow & y_{21} \quad y_{22} \quad y_{23} \quad \dots \quad y_{2n_2} \\ x_3 & \Rightarrow & y_{31} \quad y_{32} \quad y_{33} \quad \dots \quad y_{3n_3} \\ \vdots & \vdots & \vdots \quad \vdots \quad \vdots \quad \ddots \quad \vdots \\ x_m & \Rightarrow & y_{m1} \quad y_{m2} \quad y_{m3} \quad \dots \quad y_{mn_m} \end{array} \right.$$

Para entendermos a quebra, tomamos,

$$e_{ij} = y_{ij} - \hat{y}_i = (y_{ij} - \bar{y}_{i.}) + (\bar{y}_{i.} - \hat{y}_i),$$

SQ_{PE}
(Erro Puro)

SQ_{LOF}
(Falta de Ajuste)

Análise da Falta de Ajuste nos modelos de Regressão Linear Simples

Teste de Hipótese

$$\begin{cases} H_0: E(y_i) = \beta_0 + \beta_1 x_i & \text{(modelo linear adequado)} \\ H_1: E(y_i) \neq \beta_0 + \beta_1 x_i & \text{(modelo linear inadequado)} \end{cases}$$

Estatística do Teste

$$F_0 = \frac{\frac{SQ_{LOF}}{(m-2)}}{\frac{QM_{PE}}{(n-m)}} = \frac{QM_{LOF}}{QM_{PE}} \sim F_{(m-2, n-m)}$$

O valor de F_0 deve ser comparado com o quantil distribuição F de Snedecor com $(m - 2)$ graus de liberdade no numerador e $(n - m)$ graus de liberdade no denominador para um determinado nível de significância α

Fonte	GL	Soma de quadrados	Quadrado médio	F_0
Regressão	1	$SQR = \hat{\beta}_1 S_{xy}$	$QMR = \frac{SQR}{SQE}$	$F_0 = \frac{QMR}{QME}$
Resíduo	$n - 2$	$SQE = \sum_{i=1}^m \sum_{j=1}^{n_i} (y_{ij} - \hat{y}_i)^2$	$QME = \frac{SQE}{(n-2)}$	
Falta de ajuste	$m - 2$	$SQ_{LOF} = \sum_{i=1}^m n_i (\hat{y}_i - \bar{y}_{i.})^2$	$QM_{LOF} = \frac{SQ_{LOF}}{(m-2)}$	$F_0 = \frac{QM_{LOF}}{QM_{PE}}$
Erro puro	$n - m$	$SQ_{PE} = \sum_{i=1}^m \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i.})^2$	$QM_{PE} = \frac{SQ_{PE}}{(n-m)}$	
Total	$n - 1$	$SQT = S_{yy}$		

Exemplo (Motivação 1)

Em problemas de tratamento térmico deseja-se estabelecer uma relação entre a temperatura da estufa e uma característica da qualidade (Dureza). Determinar os valores de temperatura em C que “otimizam” a performance do processo de tratamento térmico em relação a estrutura metalográfica do material, avaliada em relação de dureza em HB.

Temos inicialmente

$$(y_{ij} - \hat{y}_i) = (y_{ij} - \bar{y}_{i.}) - (\bar{y}_{i.} - \hat{y}_i)$$

- $S_{xx} = \sum_{i=1}^n x_i^2 - n\bar{x}^2 = 1035750 - 20 * 227,50^2 = 625$
- $S_{yy} = \sum_{i=1}^n y_i^2 - n\bar{y}^2 = 335594 - 20 * 129,40^2 = 706,80$
- $S_{xy} = \sum_{i=1}^n x_i y_i - n\bar{x}\bar{y} = 588125 - 20 * 227,5 * 129,4 = -645$

$$SQE = \sum_{i=1}^m \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i.})^2 + \sum_{i=1}^m \sum_{j=1}^{n_i} (\bar{y}_{i.} - \hat{y}_i)^2 = 30,4 + 10,76 = 41,16$$

$$QM_{PE} = \frac{SQ_{PE}}{n - m} = \frac{30,4}{16} = 1,9$$

$$QM_{LOF} = \frac{SQ_{LOF}}{m - 2} = \frac{10,76}{2} = 5,38$$

$$F_0 = \frac{QM_{LOF}}{QM_{PE}} = \frac{5,38}{1,9} = 2,83 \sim F_{(2;16)}$$

Rejeitamos H_0 se $F_0 > F_{(2;16)}$

$$P\text{-valor} = P[F_{(2;16)} > F_0] = 0,089$$

Fonte	SQ	GL	QM	Estatística	P-valor
Regressão	665,64	1	665,64	350,3368	0
Resíduo	41,16	18	2,286		
Falta de Ajuste	10,76	2	5,38	2,8316	0,089
Erro Puro	30,4	16	1,9		
Total	706,8	19			

Análise da Falta de Ajuste nos modelos de Regressão Linear Múltipla

Teste de Hipótese

$$\begin{cases} H_0: E(y_i) = \beta_0 + \beta_1 x_i & \text{(modelo linear adequado)} \\ H_1: E(y_i) \neq \beta_0 + \beta_1 x_i & \text{(modelo linear inadequado)} \end{cases}$$

Estatística do Teste

$$F_0 = \frac{\frac{SQ_{LOF}}{(m-p-1)}}{\frac{QM_{PE}}{(n-m)}} = \frac{QM_{LOF}}{QM_{PE}} \sim F_{(m-p-1, n-m)}$$

O valor de F_0 deve ser comparado com o quantil distribuição F de Snedecor com $(m - p - 1)$ graus de liberdade no numerador e $(n - m)$ graus de liberdade no denominador para um determinado nível de significância α

Fonte	GL	Soma de quadrados	Quadrado médio	F_0
Regressão	1	$SQR = \hat{\beta}_1 S_{xy}$	$QMR = \frac{SQR}{SQE}$	$F_0 = \frac{QMR}{QME}$
Resíduo	$n - 2$	$SQE = \sum_{i=1}^m \sum_{j=1}^{n_i} (y_{ij} - \hat{y}_i)^2$	$QME = \frac{SQE}{(n-2)}$	
Falta de ajuste	$m - 2$	$SQ_{LOF} = \sum_{i=1}^m n_i (\hat{y}_i - \bar{y}_{i.})^2$	$QM_{LOF} = \frac{SQ_{LOF}}{(m-2)}$	$F_0 = \frac{QM_{LOF}}{QM_{PE}}$
Erro puro	$n - m$	$SQ_{PE} = \sum_{i=1}^m \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i.})^2$	$QM_{PE} = \frac{SQ_{PE}}{(n-m)}$	
Total	$n - 1$	$SQT = S_{yy}$		

Exemplo (Motivação 2)

O ganho de um transistor consiste na diferença entre o emissor e o coletor. A variável **Ganho** (em hFE) pode ser controlada no processo de deposição de íons por meio das variáveis **Tempo de emissão** (em minutos) e **Dose de íons** ($\times 10^{14}$).

Temos inicialmente

$$(y_{ij} - \hat{y}_i) = (y_{ij} - \bar{y}_{i.}) - (\bar{y}_{i.} - \hat{y}_i)$$

$$SQE = \sum_{i=1}^m \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i.})^2 + \sum_{i=1}^m \sum_{j=1}^{n_i} (\bar{y}_{i.} - \hat{y}_i)^2 = 2942 + 28587 = 31529$$

$$QM_{PE} = \frac{SQ_{PE}}{n - m} = \frac{2942}{15} = 196$$

$$QM_{LOF} = \frac{SQ_{LOF}}{m - 2} = \frac{28587}{12} = 2382$$

$$F_0 = \frac{QM_{LOF}}{QM_{PE}} = \frac{2382}{196} = 12,15 \sim F_{(12; 15)}$$

Rejeitamos H_0 se $F_0 > F_{(12; 15)}$

$$P\text{-valor} = P[F_{(12; 15)} > F_0] = 0$$

Fonte	SQ	GL	QM	Estatística	P-valor
Regressão	1307620	2	653810	559,89	0
Resíduo	31529	27	1168		
Falta de Ajuste	28587	12	2382	12,15	0
Erro Puro	2942	15	196		
Total	133914	29			

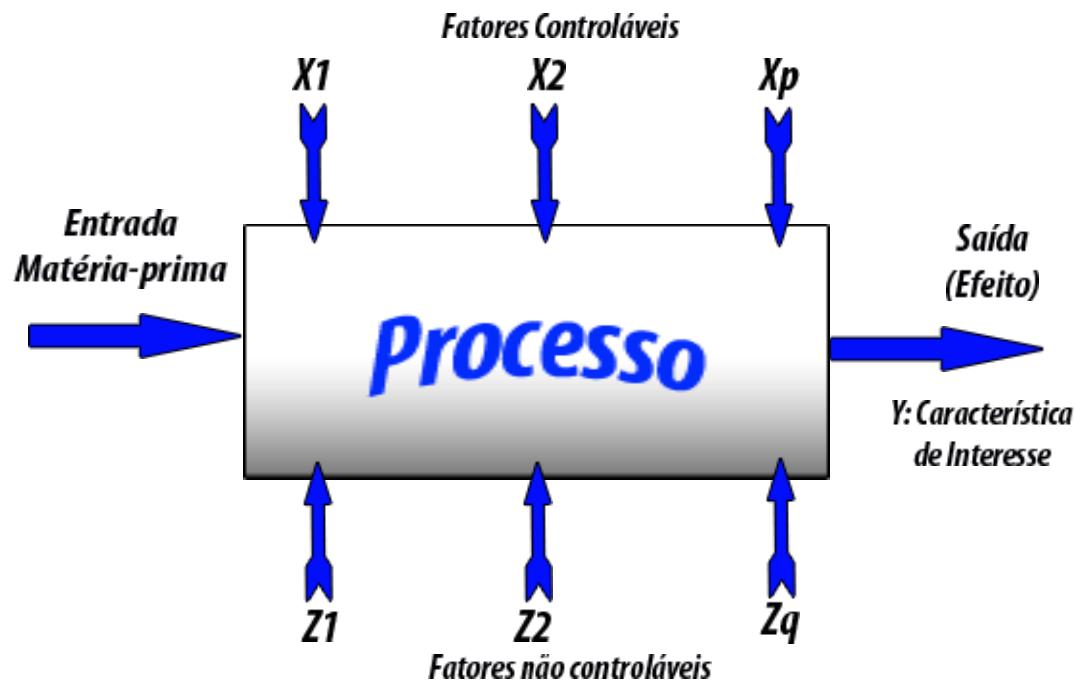
DOE - Planejamento de Experimentos

Introdução ao DOE

Experimento



É um procedimento no qual alterações propositais são feitas nas variáveis de entrada de um processo ou sistema, de modo que se possa avaliar as possíveis alterações sofridas pelas variáveis respostas, como também as razões de suas alterações.



Introdução ao DOE

Objetivos de um Experimento Planejado

3 princípios básicos do planejamento de experimentos



Redução da variação do processo e melhor concordância entre os valores obtidos e os valores pretendidos.



Redução do tempo do processo



Redução do custo operacional



Melhoria no rendimento do processo



Replicação



Aleatoriedade



Blocagem

Introdução ao DOE

Replicação

Permite a obtenção do erro experimental.

A estimativa desse erro é básica para verificar se as diferenças observadas nos dados são estatisticamente diferentes.

Se a média de uma amostra for usada para estimar o efeito de um fator no experimento, a replicação permite a obtenção de uma estimativa mais precisa desse efeito.

Aleatoriedade

Os métodos estatísticos requerem que as observações, ou os erros, sejam variáveis aleatórias distribuídas independentemente.

Os experimentos, com suas réplicas, devem ser realizados de forma aleatória, de modo a garantir a distribuição equânime de todos os fatores não considerados.

Por exemplo, ao realizarmos um experimento para determinar as variáveis determinantes do acabamento da peça em uma retífica, devemos cuidar da aleatoriedade na execução do experimento.

Blocagem

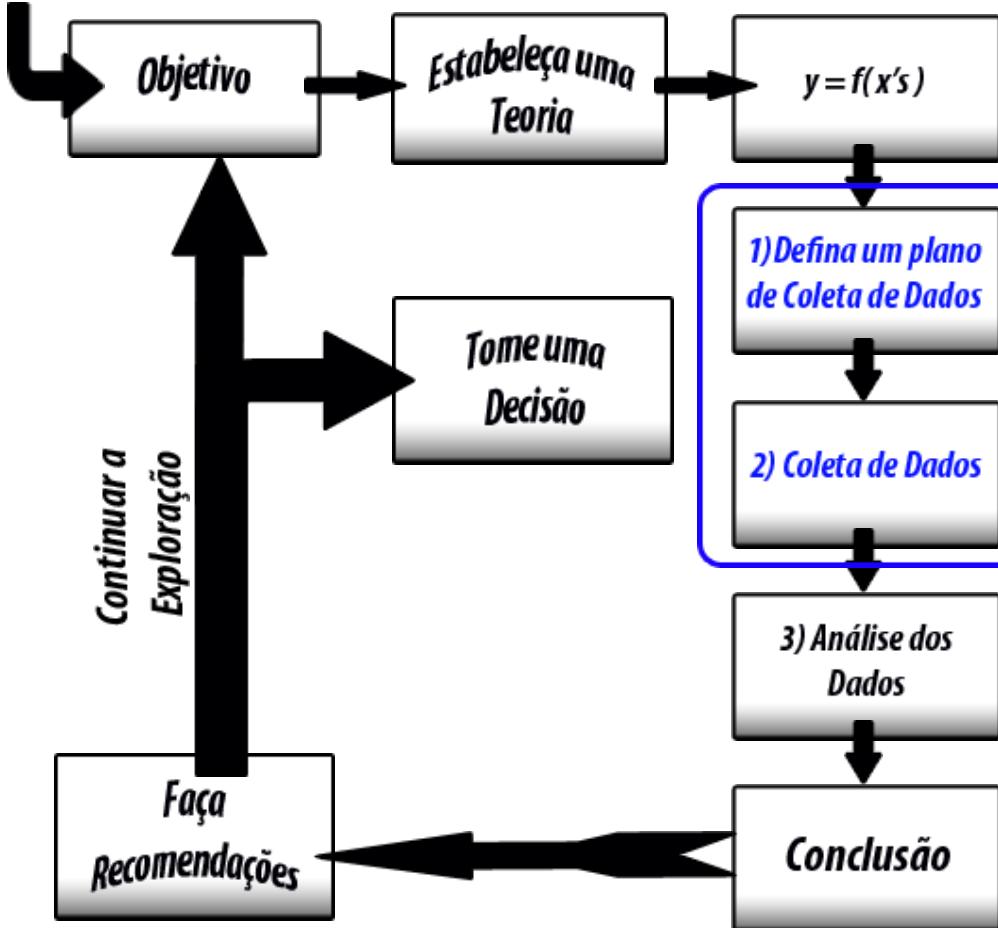
É uma técnica extremamente importante, utilizada com o objetivo de aumentar a precisão de um experimento.

Em certos processos, podemos controlar e avaliar, sistematicamente, a variabilidade resultante da presença de fatores conhecidos que perturbam o sistema, mas que não temos interesse em estudá-los.

Por exemplo, quando uma determinada medida experimental é feita por duas diferentes pessoas, levando a uma possível não homogeneidade nos dados.

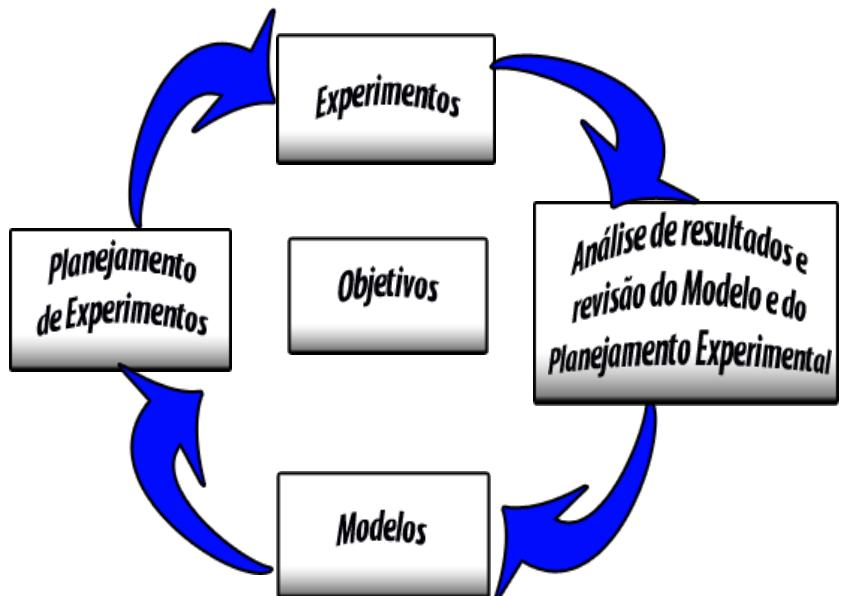
Introdução ao DOE

Fatores
Observações



Antes de começar a realizar os experimentos, os objetivos e os critérios devem estar bem claros, de modo a dar subsídios para a escolha:

- das variáveis envolvidas nos experimentos
- da faixa de variação das variáveis selecionadas
- dos níveis escolhidos para essas variáveis. No caso de muitos fatores, é melhor escolher inicialmente dois níveis
- da variável de resposta
- do planejamento experimental.



Experimentos Fatoriais Completos

Experimentos Fatoriais Completos

Experimentos
Fatoriais
Completos

São aqueles que investigam todas as possíveis combinações dos níveis dos fatores (tratamentos) em cada ensaio completo ou réplica do experimento.

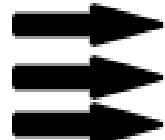
Efeito
Principal de
um Fator

É definido como a mudança na resposta produzida no nível de um fator.

Experimentos
fatoriais 2^k

Os experimentos fatoriais 2^k têm k fatores, com dois níveis cada. Um esquema de um experimento factorial pode ser visto da seguinte maneira.

Fatores



*Produto ou
Processo*



Resposta

Experimentos Fatoriais Completos

Estudar o efeito no tempo de uma determinada reação química com a variação de temperatura e concentração de um reagente.



Variável Resposta	Y : Tempo de Reação
Fatores	A : Concentração do Reagente (níveis $V_{-1} = 10\%$ e $V_{+1} = 20\%$) B : Temperatura (níveis $T_{-1} = 80^{\circ}C$ e $T_{+1} = 90^{\circ}C$)
Tratamentos	$V_{-1}T_{-1}$ - concentração em 10% e temperatura em $80^{\circ}C$ ((0)), $V_{+1}T_{-1}$ - concentração em 20% e temperatura em $80^{\circ}C$ (a), $V_{-1}T_{+1}$ - concentração em 10% e temperatura em $90^{\circ}C$ (b) $V_{+1}T_{+1}$ - concentração em 20% e temperatura em $90^{\circ}C$ (ab) (O número de tratamentos é 2^k , neste caso $2^2 = 4$)
Unidade Experimental	Período de tempo para cada reação
Réplicas	Repetições do experimento feitas sob as mesmas condições experimentais, no caso do exemplo sob o mesmo nível de temperatura e de reagente. Quanto mais réplicas, mais confiáveis os resultados do experimento.

Experimentos Fatoriais Completos

Os experimentos 2^k constituem uma importante classe de planejamento

O número de ensaios envolvidos na realização desses experimentos é relativamente pequeno.

A região experimental nestes experimentos pode ser restrita, porém pode indicar tendências e determinar uma melhor (ou mais adequada) direção para novos experimentos

Estes experimentos podem ser “aumentados” com a inclusão de novos níveis e/ou novos fatores

Os experimentos 2^k formam a base dos Experimentos Fatoriais Fracionados, onde somente alguns dos possíveis tratamentos são observados

A aleatorização é um procedimento que conduz as combinações dos níveis dos fatores em uma sequência de tratamentos de maneira que cada tratamento tenha a mesma chance de ser utilizado no experimento.

Passos para a construção de um experimento aleatório completo:

1. Enumere todas as combinações entre níveis dos fatores. De 1 a N.
2. Inclua as réplicas.
3. Gere uma sequência de números aleatórios para a sequência de 1 até N, utilizando tabelas de números aleatórios ou o computador. Conduza o experimento seguindo a sequência obtida na aleatorização

Aleatorização

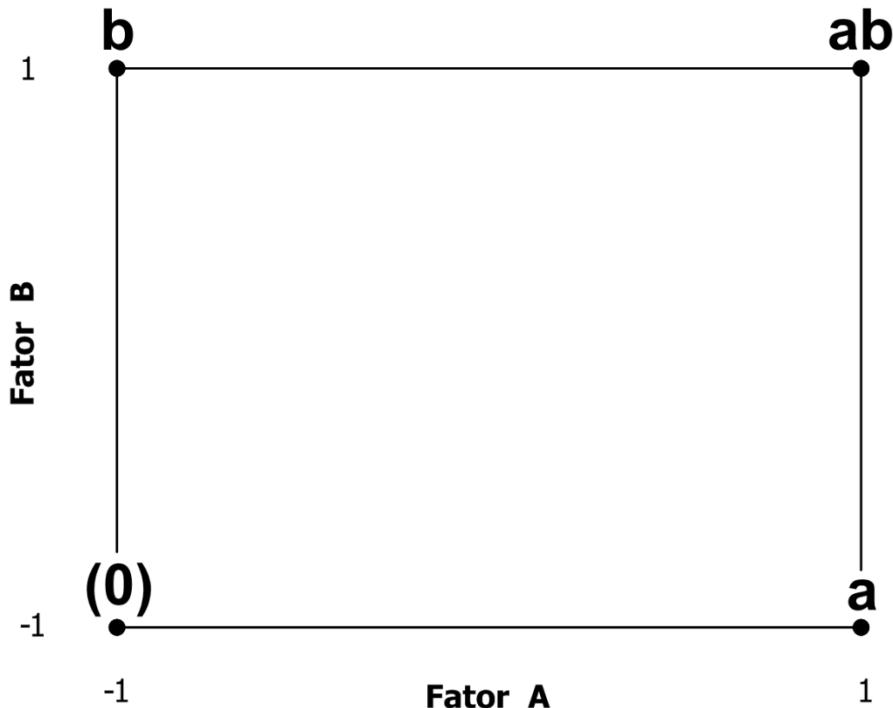
Experimentos Fatoriais Completos 2^2

Experimentos Fatoriais Completos

Notação

1. Atribua os sinais de - se nível baixo e + se nível alto para cada um dos fatores, ou
2. Atribua valor 0 ou 1 aos níveis de cada um dos fatores
3. Denotar a presença ou ausência do nível alto do fator, isto é:
 - (0) : Ambos os fatores estão ausentes (ambos em nível baixo);
 - a : Somente o fator A está presente (nível alto em A e baixo em B);
 - b : Somente o fator B está presente (nível alto em B e baixo em A);
 - ab : Ambos os fatores AB estão presentes (ambos em nível alto).

Gráfico de Médias para Y

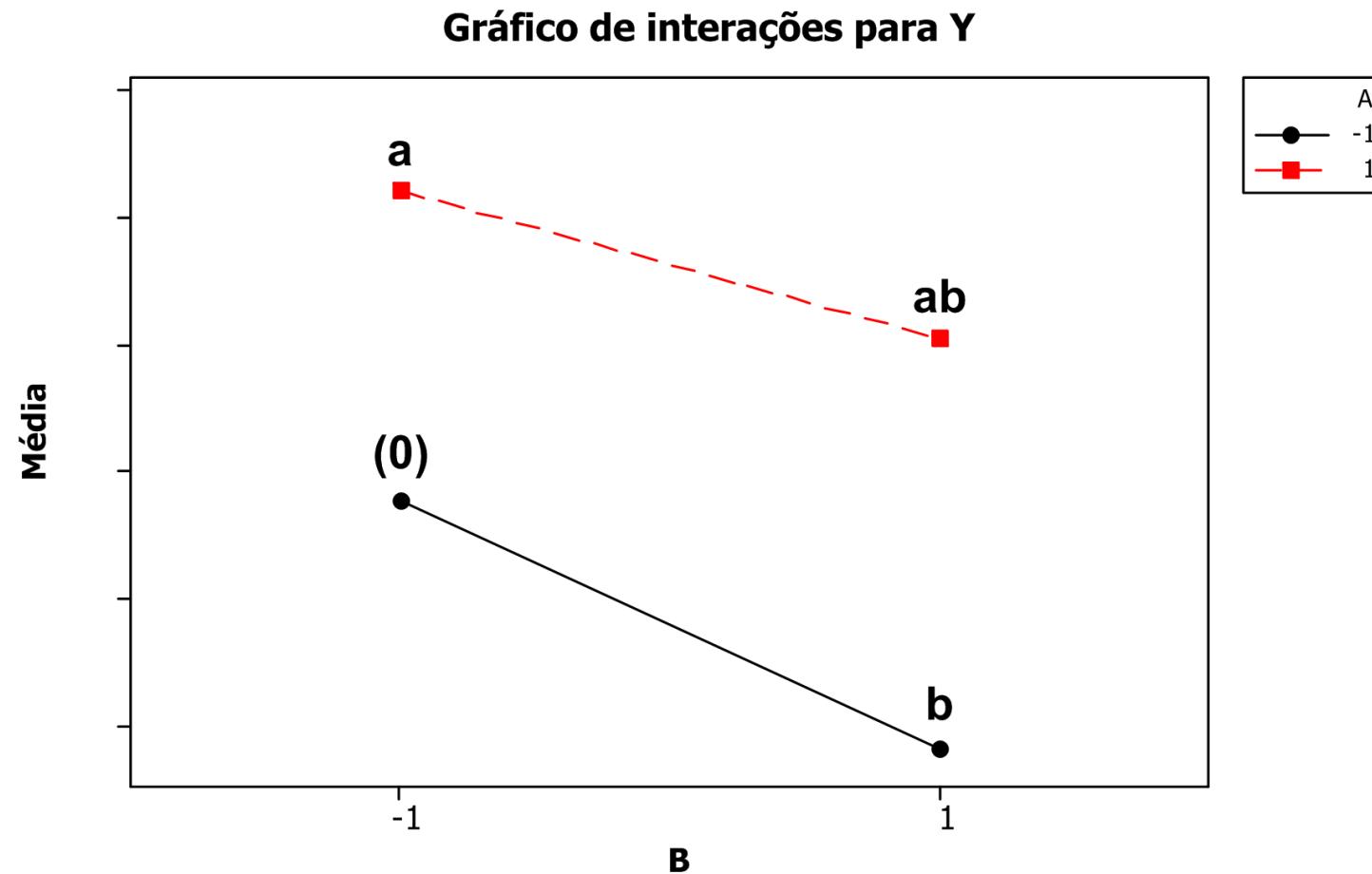


Tratamento	A	B	AB
0	-1	-1	1
a	1	-1	-1
b	-1	1	-1
ab	1	1	1

Experimentos Fatoriais Completos

Interações

A interação ocorre entre dois ou mais fatores se o efeito de um fator na resposta depende do nível dos outros fatores.



Exemplo

Exemplo 3.1.1: Estudar o efeito no tempo de uma determinada reação química com a variação de temperatura e concentração de um reagente.

Variável Resposta	Y : Tempo de Reação
Fatores	A: Concentração do Reagente (níveis $V_{-1} = 10\%$ e $V_{+1} = 20\%$) B: Temperatura (níveis $T_{-1} = 80^\circ C$ e $T_{+1} = 90^\circ C$)
Tratamentos	$V_{-1}T_{-1}$ - concentração em 10% e temperatura em $80^\circ C$ ((0)), $V_{+1}T_{-1}$ - concentração em 20% e temperatura em $80^\circ C$ (a), $V_{-1}T_{+1}$ - concentração em 10% e temperatura em $90^\circ C$ (b) $V_{+1}T_{+1}$ - concentração em 20% e temperatura em $90^\circ C$ (ab) (O número de tratamentos é 2^k , neste caso $2^2 = 4$)
Unidade Experimental	Período de tempo para cada reação
Réplicas	Repetições do experimento feitas sob as mesmas condições experimentais, no caso do exemplo sob o mesmo nível de temperatura e de reagente. Quanto mais réplicas, mais confiáveis os resultados do experimento.

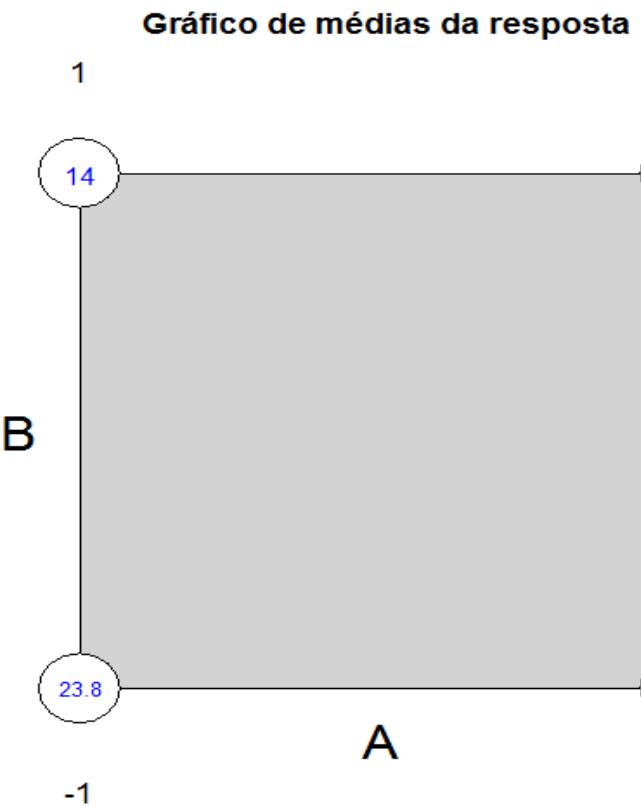
Considere 3 réplicas deste experimento (completo), com os dados abaixo, e a ordem de execução entre parênteses:

Tratamento	A	B	AB	y_1	y_2	y_3	\bar{y}
(0)	-1	-1	+1	26,6(1)	22,0(7)	22,8(10)	23,8
(a)	+1	-1	-1	40,9(4)	36,4(9)	36,7(12)	38
(b)	-1	+1	-1	11,8(3)	15,9(8)	14,3(11)	14
(ab)	+1	+1	+1	34,0(2)	29,0(5)	33,6(6)	32,2

Exemplo

Exemplo 3.1.1: Estudar o efeito no tempo de uma determinada reação química com a variação de temperatura e concentração de um reagente.

Primeiramente fazemos um gráfico com as médias dos valores de Y para os diferentes tratamentos:



Observamos que a melhor configuração seria A_- e B_+ . No entanto, a interação entre A e B pode ser significativa, assim nada podemos concluir neste momento.

- \bar{Y}_{A+} a média de Y com o fator A no nível + (alto);
- \bar{Y}_{A-} a média de Y com o fator A no nível - (baixo);
- e assim para os demais fatores.

Calculamos os efeitos médios dos fatores A, B como sendo:

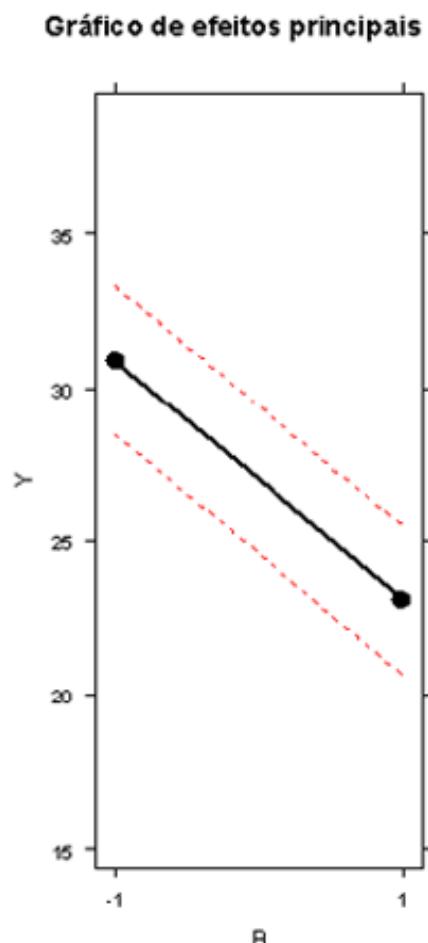
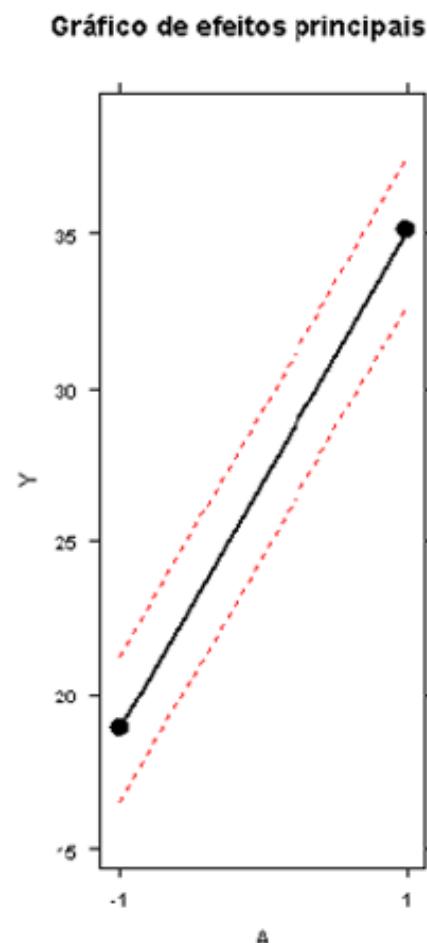
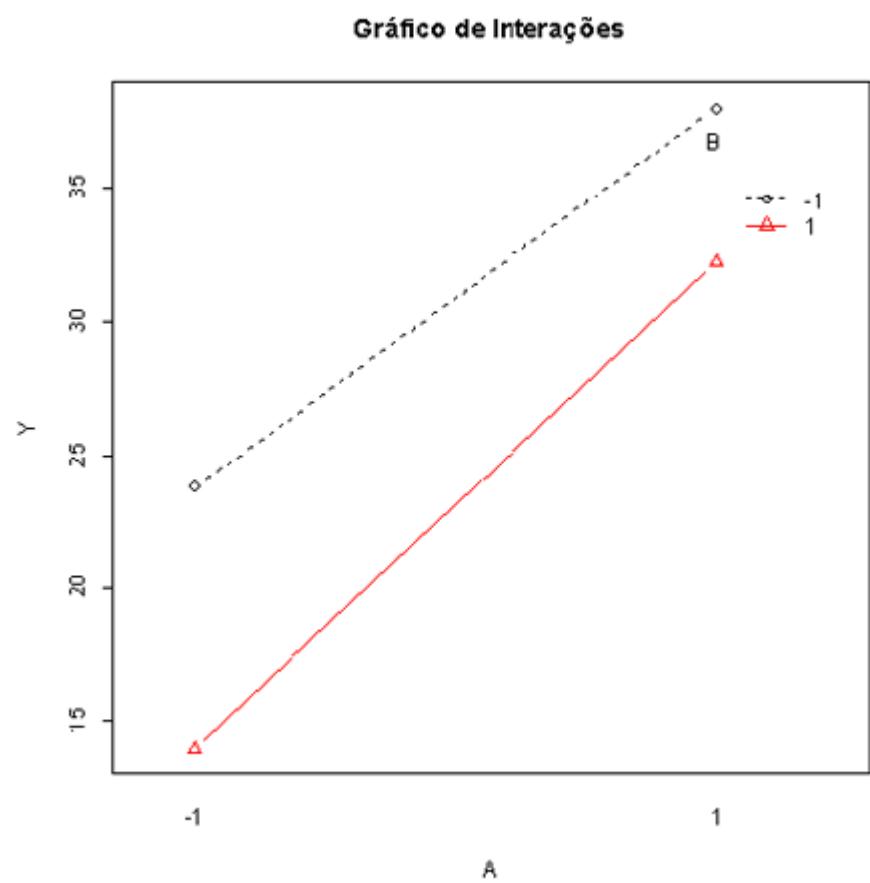
$$A = \bar{Y}_{A+} - \bar{Y}_{A-} = \frac{ab + a}{2} - \frac{b + (0)}{2} = \frac{32,2 + 38}{2} - \frac{14 + 23,8}{2} = 16,2$$
$$B = \bar{Y}_{B+} - \bar{Y}_{B-} = \frac{ab + b}{2} - \frac{a + (0)}{2} = \frac{32,2 + 14}{2} - \frac{38 + 23,8}{2} = -7,8$$

Efeito da interação entre os fatores A e B

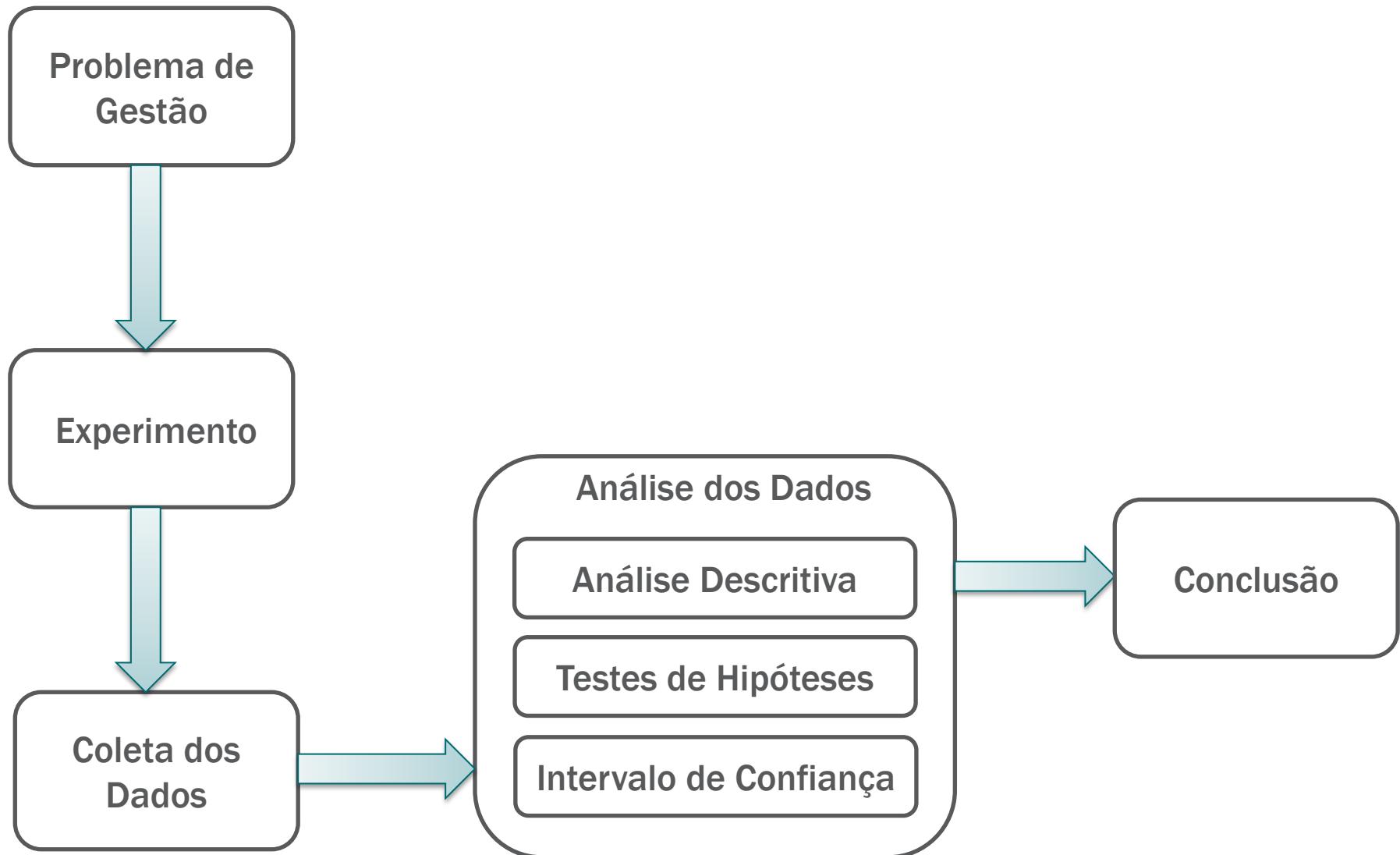
$$AB = \frac{ab - a}{2} - \frac{b - (0)}{2} = \frac{32,2 - 38}{2} - \frac{14 - 23,8}{2} = 2$$

Exemplo

Exemplo 3.1.1: Estudar o efeito no tempo de uma determinada reação química com a variação de temperatura e concentração de um reagente.

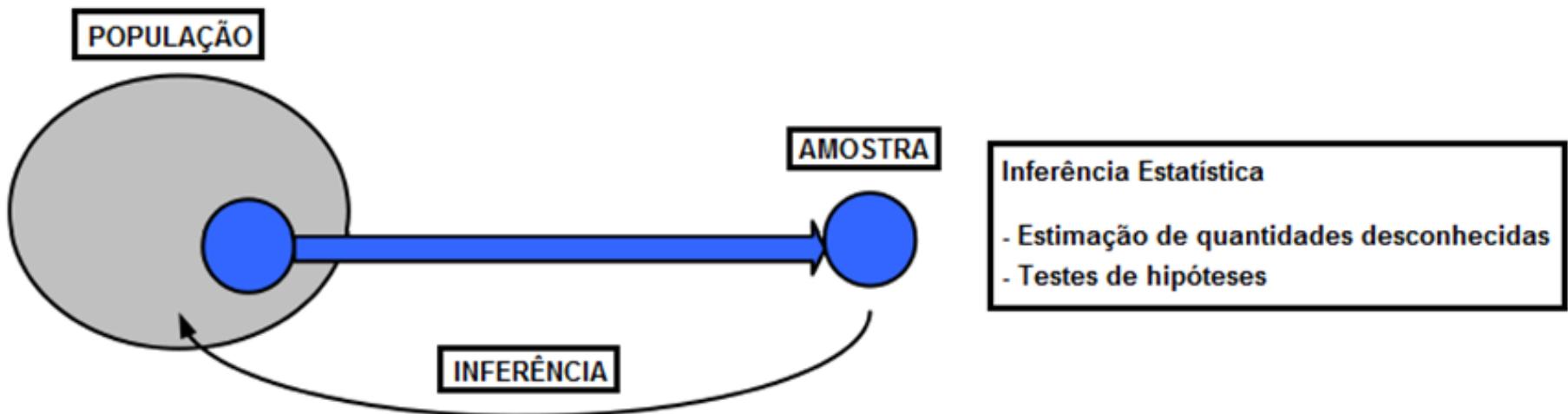


Problema de Gestão



População: Conjunto de elementos com uma mesma característica em comum.

Amostra: Subconjunto de elementos de uma população representativos para estudar a característica de interesse.



Inferência

Variável Aleatória: Uma variável aleatória é um característico numérico do resultado de um experimento, isto é, uma função que associa aos eventos de um experimento, um número real.

$$X : \Omega \rightarrow R$$

Exemplo: Considere o lançamento de uma moeda. Os resultados possíveis deste experimento são cara ou coroa.

Neste caso, podemos definir $\Omega = \{\text{cara}, \text{coroa}\}$ e podemos associar a seguinte variável aleatória

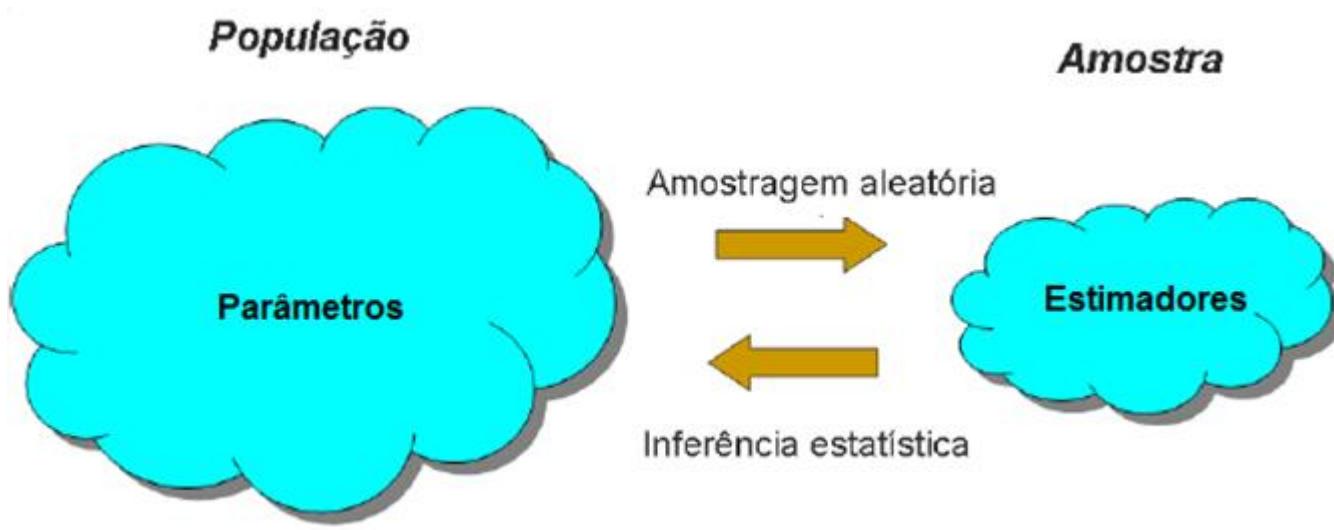
$$X(\{\text{cara}\}) = 1$$

$$X(\{\text{coroa}\}) = 0$$

Parâmetro: Característica numérica de interesse (desconhecida) da distribuição dos elementos de uma população.

Estimador: Elemento construído a partir de uma amostra com a finalidade de representar ou estimar um parâmetro de interesse da população.

Estimativa: Valor numérico que um estimador assume.



Exemplo:

Para saber qual é a altura média do brasileiro utilizamos uma amostra para termos uma estimativa deste valor, pois é completamente inviável medir a altura de todos os brasileiros para calcular sua média.

Testes de Hipóteses

**Por que precisamos
deles?**

**Para conseguir evidências estatísticas para
a tomada de decisões.**

Matriz de Decisão

Use o teste apropriado para avaliar os seus dados

		Variável X	
		Discreta	Contínua
Variável Y	Discreta	Teste de Proporção Tabela Cruzada Regressão Logística Regressão Poisson	Regressão Logística Regressão Poisson
	Contínua	ANOVA Teste de médias Teste de medianas Teste de Variâncias	Correlação Regressão

Análise dos efeitos via ANOVA

ANOVA

Ferramenta para análise do comportamento de diversos tratamentos para dois fatores, aplicados ao processo e/ou produto.

O comportamento é avaliado com relação a uma resposta do processo e/ou produto.

Sejam A o primeiro fator, com a níveis e B o segundo fator, com b níveis e r as réplicas do experimento.

Fator A	Fator B				Média
	1	2	...	b	
1	y_{111}, \dots, y_{11r}	y_{121}, \dots, y_{12r}	...	y_{1b1}, \dots, y_{1br}	$\bar{y}_{1..}$
2	y_{211}, \dots, y_{21r}	y_{221}, \dots, y_{22r}	...	y_{2b1}, \dots, y_{2br}	$\bar{y}_{2..}$
:	:	:	:	:	:
a	y_{a11}, \dots, y_{a1r}	y_{a21}, \dots, y_{a2r}	...	y_{ab1}, \dots, y_{abr}	$\bar{y}_{a..}$
Média	$\bar{y}_{.1.}$	$\bar{y}_{.2.}$...	$\bar{y}_{.b.}$	$\bar{y}_{...}$

Análise dos efeitos via ANOVA

O modelo com dois fatores balanceados e com efeitos cruzados com interação é dado por:

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \tau_{ij} + \varepsilon_{ijk},$$

com,

$i = 1, 2, \dots, a$ Fator A

$j = 1, 2, \dots, b$ Fator B

$k = 1, \dots, r$ Réplica

Em que

- μ : média geral dos dados
- α_i : efeito do nível i do fator A
- β_j : efeito do nível j do fator B
- τ_{ij} : efeito da interação ij entre os fatores
- ε_{ijk} : componente aleatória do erro

Notações:

- $y_{i..} = \sum_{j=1}^b \sum_{k=1}^r y_{ijk}$ = soma das observações do nível i do fator A
- $\bar{y}_{i..} = \frac{y_{i..}}{br}$ = média das observações do nível i do fator A
- $y_{...} = \sum_{i=1}^n \sum_{j=1}^{n_i} y_{ij}$ = soma de todas as observações
- $\bar{y}_{...} = \frac{1}{b a r} \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^r y_{ijk}$ = média geral das observações
- Analogamente, $\bar{y}_{.j.} = \frac{1}{a r} \sum_{i=1}^a \sum_{k=1}^r y_{ijk}$, $\bar{y}_{i.j.} = \frac{1}{r} \sum_{k=1}^r y_{ijk}$

Análise dos efeitos via ANOVA

Hipóteses

$\varepsilon_{ijk} \sim N(0, \sigma^2)$ independentes.

$y_{ijk} \sim N(\mu + \alpha_i + \beta_j + \tau_{ij}, \sigma^2)$ independentes para $i = 1, \dots, k$.

Teste de Hipóteses

Efeito do Fator A
 $H_0: \alpha_1 = \alpha_2 = \dots = \alpha_a$
 $H_1:$ pelo menos um é diferente.

Efeito do Fator B
 $H_0: \beta_1 = \beta_2 = \dots = \beta_b$
 $H_1:$ pelo menos um é diferente.

Efeito da Interação $A \times B$
 $H_0: \tau_{ij} = 0$ para todos os valores de i e j
 $H_1: \tau_{ij} \neq 0$

Quebra da Soma de Quadrados

$$SQ_T = SQ_A + SQ_B + SQ_{AB} + SQ_E$$

Onde:

$$SQ_A = br \sum_{i=1}^a (\bar{y}_{i..} - \bar{y}_{...})^2 = \frac{1}{br} \sum_{i=1}^a y_{i..}^2 - \frac{\bar{y}_{...}^2}{abr}$$

$$SQ_B = ar \sum_{i=1}^b (\bar{y}_{.j.} - \bar{y}_{...})^2 = \frac{1}{ar} \sum_{i=1}^b y_{.j.}^2 - \frac{\bar{y}_{...}^2}{abr}$$

E a soma de quadrados do efeito da interação é:

$$SQ_{AB} = r \sum_{i=1}^a \sum_{j=1}^b (\bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}_{...})^2 = \frac{1}{r} \sum_{i=1}^a \sum_{j=1}^b y_{ij.}^2 - \frac{\bar{y}_{...}^2}{abr} - SQ_A - SQ_B$$

Temos também que a Soma de Quadrados Total é dada por:

$$SQ_T = \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^r (y_{ijk} - \bar{y}_{...})^2 = \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^r y_{ijk}^2 - \frac{\bar{y}_{...}^2}{abr}$$

e finalmente a Soma de Quadrados dos Erros:

$$SQ_E = \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^r (y_{ijk} - \bar{y}_{ij.})^2 = \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^r y_{ijk}^2 - \frac{1}{r} \sum_{i=1}^a \sum_{j=1}^b y_{ij.}^2$$

Como estamos trabalhando com experimentos fatoriais 2^2 então tomando $a = b = 2$, temos

$$SQ_A = br \sum_{i=1}^a (\bar{y}_{i..} - \bar{y}_{...})^2 = 2r[(\bar{y}_{2..} - \bar{y}_{...}) + (\bar{y}_{1..} - \bar{y}_{...})]^2$$

Ou seja,

$$\begin{aligned} SQ_A &= 2r \left[\left(\bar{y}_{2..} - \frac{\bar{y}_{2..} - \bar{y}_{1..}}{2} \right) + \left(\bar{y}_{1..} - \frac{\bar{y}_{2..} - \bar{y}_{1..}}{2} \right) \right]^2 \\ &= 2r \left[\left(\frac{\bar{y}_{2..} - \bar{y}_{1..}}{2} \right) + \left(\frac{\bar{y}_{1..} - \bar{y}_{2..}}{2} \right) \right]^2 \end{aligned}$$

Então

$$SQ_A = 4r \left[\left(\frac{\bar{y}_{2..} - \bar{y}_{1..}}{2} \right) \right]^2 = r(\bar{y}_{2..} - \bar{y}_{1..})^2 = r(\text{efeito de } A)^2$$

Para os experimentos fatoriais 2^2 , tem-se

$$SQ_A = r(\text{efeito de } A)^2$$

$$SQ_B = r(\text{efeito de } B)^2$$

$$SQ_{AB} = r(\text{efeito de } AB)^2$$

Assim, constrói-se a Tabela da Anova.

Análise dos efeitos via ANOVA

A soma de quadrados total é definida como medida da variabilidade total dos dados



$$SQA = r(\bar{y}_{2..} - \bar{y}_{1..})^2 = r(\text{efeito de } A)^2$$

$$SQB = r(\text{efeito de } B)^2$$

$$SQAB = r(\text{efeito de } AB)^2$$

Fator	Soma de Quadrados	Graus de Liberdade	Quadrados Médios	F_0
Fator A	SQA	$a - 1$	$QMA = \frac{SQA}{a - 1}$	$F_A = \frac{QMA}{QME}$
Fator B	SQB	$b - 1$	$QMB = \frac{SQB}{b - 1}$	$F_B = \frac{QMB}{QME}$
Interação ($A \times B$)	SQAB	$(a - 1)(b - 1)$	$QMAB = \frac{SQAB}{(a - 1)(b - 1)}$	$F_{AB} = \frac{QMAB}{QME}$
Erro	SQE	$ab(r - 1)$	$QME = \frac{SQE}{ab(r - 1)}$	
Total	SQT	$abr - 1$		

Exemplo

Exemplo 3.1.1: Estudar o efeito no tempo de uma determinada reação química com a variação de temperatura e concentração de um reagente.

ANOVA					
DADOS DO PROCESSO					
Tabela da Anova					
A	1	787,32	787,32	129,2807882	0
B	1	182,52	182,52	29,97044335	0,0006
A:B	1	12	12	1,97044335	0,198
Resíduos	8	48,72	6,09		

- Para o fator A, $F_0 = 129,28 > F_{(0,95 ;1 ;8)} = 5,32$; portanto o fator A é significativo.
- Para o fator B, $F_0 = 29,97 > F_{(0,95 ;1 ;8)} = 5,32$; então o fator B também é significativo.
- Mas a interação não é significativa, pois $F_0 = 1,97 < F_{(0,95 ;1 ;8)} = 5,32$.

Vamos observar as mudanças quando passamos de um nível para o outro:

$$\begin{aligned}\square \quad \bar{Y}_{1..} &= \frac{(26,6+11,8+22+15,9+22,8+14,3)}{6} = 18,9 \\ \square \quad \bar{Y}_{2..} &= \frac{(40,9+34+36,4+29+36,7+33,6)}{6} = 35,1 \\ \square \quad \bar{Y}_{.1.} &= \frac{(26,6+40,9+22+36,4+22,8+36,7)}{6} = 30,9 \\ \square \quad \bar{Y}_{.2.} &= \frac{(11,8+34+15,9+29+14,3+33,6)}{6} = 23,1\end{aligned}$$

Portanto, a melhor configuração para se obter o menor tempo de reação é A_- e B_+ , ou seja, a concentração em 10% e a temperatura em 90°C.

Análise de Resíduos - ANOVA

Hipóteses

- $\varepsilon_{ij} \sim N(0, \sigma^2)$ independentes e identicamente distribuídos;
- Modelo para os dados:

$$y_{ijk} = \mu + \alpha_i + \beta_j + \tau_{ij} + \varepsilon_{ijk}$$

Na prática, precisamos verificar se estas suposições são válidas. Violações nestas suposições são verificadas através dos resíduos.

O resíduo para a j-ésima observação do nível i

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

Em que \hat{y}_{ijk} é uma estimativa da observação y_{ijk}

$$\hat{y}_{ijk} = \bar{y}_{...} + (\bar{y}_{i..} - \bar{y}_{...}) = \hat{\mu} + \hat{\alpha}_i + \hat{\beta}_j = \bar{y}_{i..}$$

Se o modelo for adequado, os resíduos observados devem refletir essas propriedades.

Propriedades dos resíduos

Média



$$\bar{e} = \frac{\sum_{i=1}^n e_i}{n} = 0$$

Variância

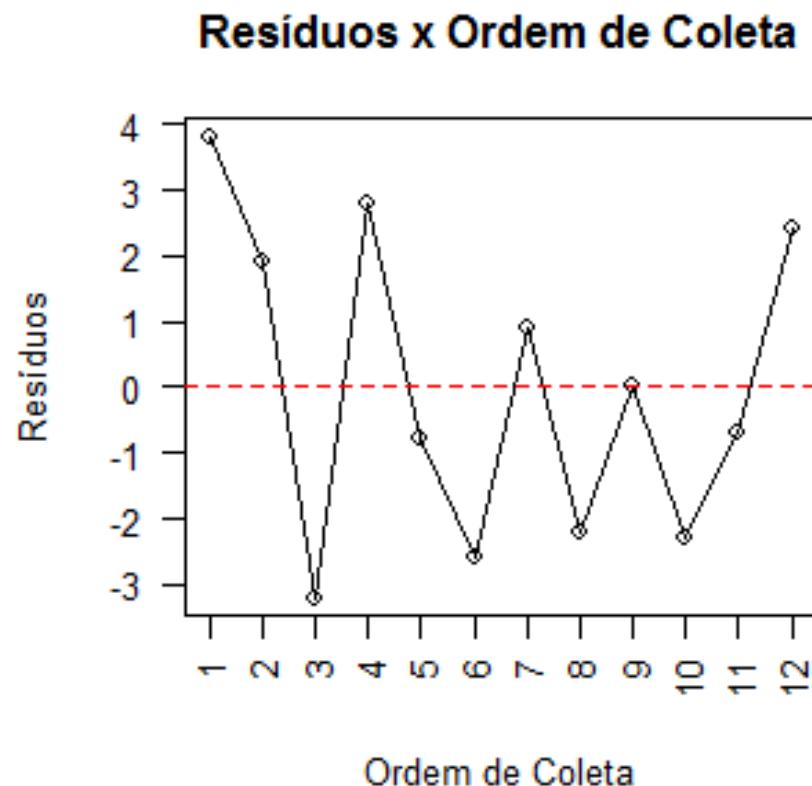


$$\frac{\sum_{i=1}^n (e_i - \bar{e})^2}{n-2} = \frac{\sum_{i=1}^n e_i^2}{n-2} = \frac{SQE}{n-2} = QME$$

Exemplo

Exemplo 3.1.1: Estudar o efeito no tempo de uma determinada reação química com a variação de temperatura e concentração de um reagente.

Resíduos
3,8
1,9
-3,2
2,8
-0,8
-2,6
0,9
-2,2
3,17948E-15
-2,3
-0,7
2,4



Análise de Resíduos

Grande parte dos problemas que encontramos na prática, são solucionados, considerando algumas suposições iniciais



Assumir a normalidade dos dados é o primeiro passo que tomamos para simplificar sua análise. Para dar suporte a esta suposição, consideramos os testes:

- Teste de Anderson-Darling;
- Teste Kolmogorov-Smirnov;
- Teste Shapiro-Wilk.

E o gráfico “Papel de Probabilidade”

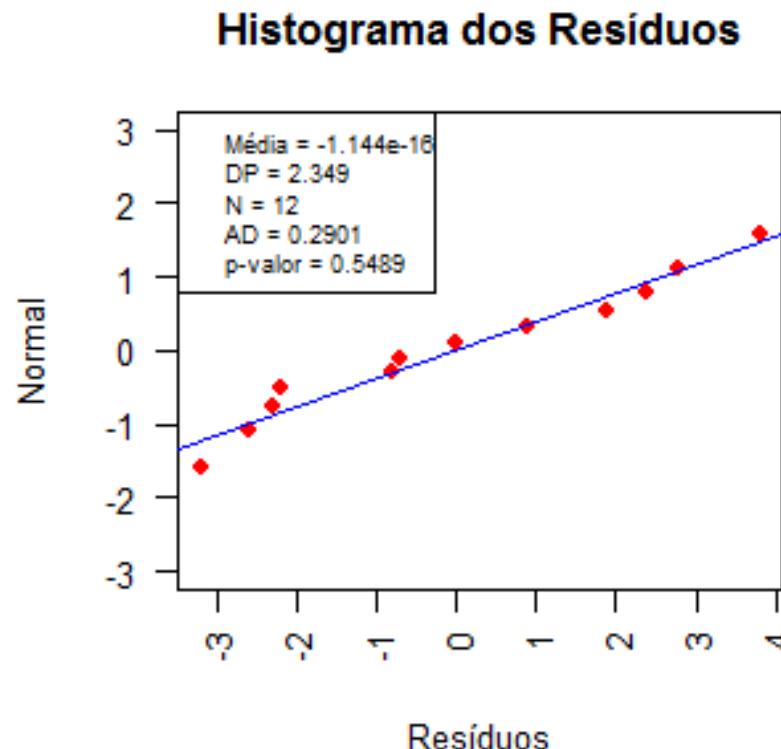
Papel de probabilidade e Teste de Anderson-Darling

Avaliamos a normalidade dos resíduos através do gráfico “papel de probabilidade” e do teste de Anderson-Darling. Consideramos a seguinte hipótese:

H_0 : os resíduos seguem uma distribuição Normal

H_1 : os resíduos Não seguem uma distribuição Normal

Para o exemplo, como o P-valor é alto (aproximadamente 0,54) não rejeitamos a hipótese de normalidade dos resíduos.



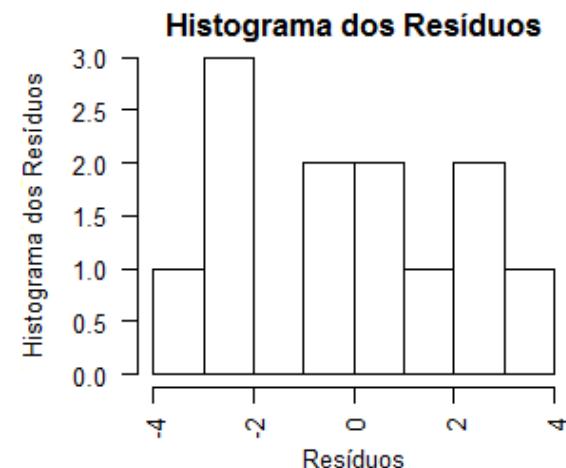
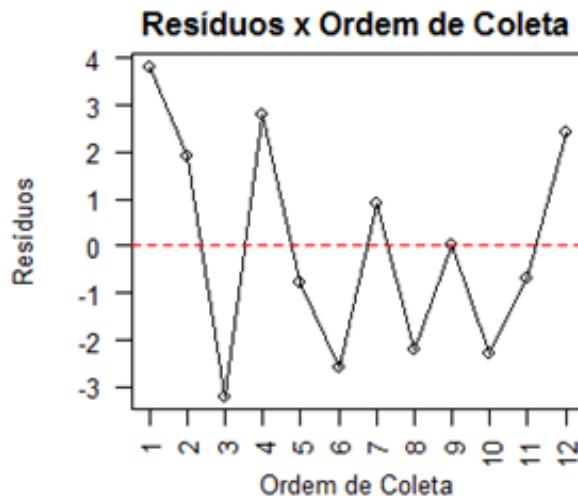
Papel de probabilidade e Teste de Anderson-Darling

Histograma dos resíduos

Também desenvolvemos o histograma dos resíduos para avaliar sua dispersão e distribuição.

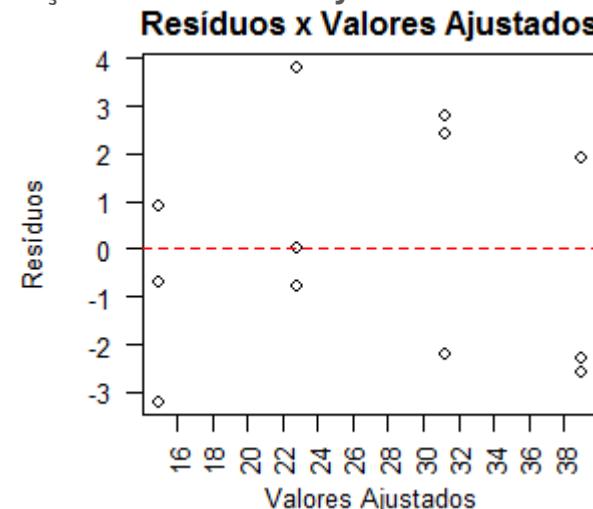
Resíduos versus a ordem de coleta dos dados

Com esse gráfico obtemos indícios da independência ou não entre os resíduos. Se algum comportamento sistemático for observado no gráfico, temos indícios de que alguma variável "extra" influenciou nos resultados do experimento, fato que viola uma das premissas básicas da ANOVA e compromete nossas conclusões.



Resíduos versus valores ajustados

Com esse gráfico temos indícios sobre o comportamento da variância dos resíduos com relação aos valores ajustados.



Análise dos efeitos via regressão linear

Modelo



$$Y = \beta_0 + \beta_1 X_i + \beta_2 X_j + \beta_{12} X_i X_j + \varepsilon$$

Em que

- β_0 : é a média geral da resposta
- X_i : assume valor -1 ou 1, dependendo do nível do fator A
- X_j : assume valor -1 ou 1, dependendo do nível do fator B
- $X_{ij} = X_i X_j$

Estimativa dos parâmetros do Modelo de Regressão Linear

- método de mínimos quadrados é o mais utilizado para estimar os parâmetros do modelo de regressão linear.
- considera-se p variáveis explicativas x, p + 1 parâmetros do modelo e n observações, com n > p.

Os dados de uma regressão linear podem ser representados da seguinte forma:

y	x ₁	x ₂	...	x _p
y ₁	x ₁₁	x ₁₂	...	x _{1p}
y ₂	x ₂₁	x ₂₂	...	x _{2p}
:	:	:	..,	:
y _n	x _{n1}	x _{n2}	...	x _{np}

Análise dos efeitos via regressão linear

Hipóteses

ε é uma variável aleatória tal que:

$$E(\varepsilon) = 0 \text{ e } Var(\varepsilon) = \sigma^2$$

supõe-se que os erros experimentais i são não correlacionados.

Modelo de regressão linear pode ser representado na seguinte forma matricial

$$Y = X\beta + \varepsilon$$

Em que

- $Y = (y_1 \ y_2 \ \dots \ y_n)^T$
- $\beta = (\beta_1 \ \beta_2 \ \dots \ \beta_n)^T$
- $\varepsilon = (\varepsilon_1 \ \varepsilon_2 \ \dots \ \varepsilon_n)^T$

A matriz X é definida por:

$$X = \begin{pmatrix} 1 & x_{11} & x_{12} & \dots & x_{1k} \\ 1 & x_{21} & x_{22} & \dots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{nk} \end{pmatrix}$$

Vamos encontrar os estimadores que minimizam o erro quadrático

$$E = \sum_j (\varepsilon_j)^2 = (Y - X\beta)^T(Y - X\beta) = Y^T Y - 2\beta^T X^T Y + \beta^T X^T X\beta$$

Calculando as derivadas de E em relação a β e igualar a zero, obtemos

$$\frac{\partial E}{\partial \beta} = -2X^T Y - 2X^T X\beta = 0$$

Substituindo β por $\hat{\beta}$, obtemos

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

O modelo ajustado é definido por

$$\hat{Y} = X\hat{\beta}$$

Vetor de Resíduos

$$e = Y - \hat{Y}$$

Análise dos efeitos via regressão linear

Experimento factorial 2^2

Y	I	X_1	X_2	$X_3 = X_1 X_2$
(0)	1	-1	-1	+1
a	1	+1	-1	-1
b	1	-1	+1	-1
ab	1	+1	+1	+1

Matricialmente, podemos expressar o modelo da seguinte forma:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \varepsilon$$

Em que

$$\mathbf{Y} = \begin{pmatrix} (0) \\ a \\ b \\ ab \end{pmatrix}, \mathbf{X} = \begin{pmatrix} 1 & -1 & -1 & +1 \\ 1 & +1 & -1 & -1 \\ 1 & -1 & +1 & -1 \\ 1 & +1 & +1 & +1 \end{pmatrix}, \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix}, \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \end{pmatrix}$$

A seguir, vamos encontrar os estimadores de mínimos quadrados

$$\hat{\boldsymbol{\beta}} = \left[\begin{pmatrix} 1 & 1 & 1 & 1 \\ -1 & +1 & -1 & +1 \\ -1 & -1 & +1 & +1 \\ +1 & -1 & -1 & +1 \end{pmatrix} \begin{pmatrix} 1 & -1 & -1 & +1 \\ 1 & +1 & -1 & -1 \\ 1 & -1 & +1 & -1 \\ 1 & +1 & +1 & +1 \end{pmatrix} \right]^{-1} \begin{pmatrix} 1 & 1 & 1 & 1 \\ -1 & +1 & -1 & +1 \\ -1 & -1 & +1 & +1 \\ +1 & -1 & -1 & +1 \end{pmatrix} \begin{pmatrix} (0) \\ a \\ b \\ ab \end{pmatrix} = \begin{pmatrix} 1/4 & 0 & 0 & 0 \\ 0 & 1/4 & 0 & 0 \\ 0 & 0 & 1/4 & 0 \\ 0 & 0 & 0 & 1/4 \end{pmatrix} \begin{pmatrix} (0) + a + b + ab \\ -(0) + a - b + ab \\ -(0) - a + b + ab \\ (0) - a - b + ab \end{pmatrix}$$

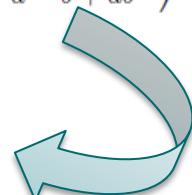
$$\hat{\beta}_0 = \frac{(0) + a + b + ab}{4}$$

$$\hat{\beta}_1 = \frac{-(0) + a - b + ab}{4} = \frac{\text{efeito de } A}{2}$$

$$\hat{\beta}_2 = \frac{-(0) - a + b + ab}{4} = \frac{\text{efeito de } B}{2}$$

$$\hat{\beta}_3 = \frac{(0) - a - b + ab}{4} = \frac{\text{efeito de } AB}{2}$$

Observe que os estimadores $\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3$ correspondem ao efeito do fator A dividido por 2, efeito do fator B dividido por 2 e efeito da interação AB dividido por 2, respectivamente.



Inferência para os coeficientes da regressão

As hipóteses para testar a significância de qualquer coeficiente de regressão é dado por:

$$H_0: \beta_j = 0$$

$$H_1: \beta_j \neq 0$$

$$j = 0, 1, \dots, p$$

A estatística do teste é dada por:

$$T_0 = \frac{\hat{\beta}_j - \beta_j}{\sqrt{\hat{\sigma}^2 C_{jj}}} \sim t_{(n-p-1)} \text{ (Sob } H_0)$$

Onde:

C_{jj} é um elemento da diagonal da matriz $(X'X)^{-1}$ correspondente a $\hat{\beta}_j$. No caso do experimento factorial 2^2 , $C_{jj} = 4r$, no qual r é o número de réplicas .

$$\hat{\sigma}^2 = \frac{SQ_E}{n - k - 1}$$

O valor de T_0 deve ser comparado com o quantil distribuição t- Student com $n-p-1$ graus de liberdade para um determinado nível de significância α

$$P - valor = 2 * P(t_{(n-p-1)} > |T_0|)$$

Exemplo

Exemplo 3.1.1: Estudar o efeito no tempo de uma determinada reação química com a variação de temperatura e concentração de um reagente.

Tratamento	A	B	AB	Y_1	Y_2	Y_3	\bar{Y}
(0)	-1	-1	+1	26,6(1)	22,0(7)	22,8(10)	23,8
(a)	+1	-1	-1	40,9(4)	36,4(9)	36,7(12)	38
(b)	-1	+1	-1	11,8(3)	15,9(8)	14,3(11)	14
(ab)	+1	+1	+1	34,0(2)	29,0(5)	33,6(6)	32,2

$$Y = \begin{pmatrix} 26,6 \\ 40,9 \\ 11,8 \\ 34 \\ 22 \\ 36,4 \\ 15,9 \\ 29 \\ 22,8 \\ 36,7 \\ 14,3 \\ 33,6 \end{pmatrix}, X = \begin{pmatrix} 1 & (A) & (B) & (AB) \\ 1 & -1 & -1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & 1 & 1 \\ 1 & -1 & -1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & 1 & 1 \\ 1 & -1 & -1 & 1 \\ 1 & 1 & 1 & -1 \\ 1 & -1 & 1 & -1 \end{pmatrix} \quad \Rightarrow \quad (X'X)^{-1} = \begin{pmatrix} \frac{1}{12} & 0 & 0 & 0 \\ 0 & \frac{1}{12} & 0 & 0 \\ 0 & 0 & \frac{1}{12} & 0 \\ 0 & 0 & 0 & \frac{1}{12} \end{pmatrix} \quad X'Y = \begin{pmatrix} 324 \\ 97,2 \\ -46,8 \\ 12 \end{pmatrix}$$

$$\hat{\beta} = (X'X)^{-1}X'Y = \begin{pmatrix} 27 \\ 8,1 \\ -3,9 \\ 1 \end{pmatrix}$$

Modelo Ajustado

$$Y = 27 + 8,1X_1 - 3,9X_2 + 1X_3$$

$$\hat{\beta}_0 = 27$$

$$\hat{\beta}_1 = \frac{\text{efeito de A}}{2} = 8,1$$

$$\hat{\beta}_2 = \frac{\text{efeito de B}}{2} = -3,9$$

$$\hat{\beta}_3 = \frac{\text{efeito de AB}}{2} = 1$$

Exemplo

Estudar o efeito no tempo de uma determinada reação química com a variação de temperatura e concentração de um reagente.

As hipóteses para testar a significância de qualquer coeficiente de regressão é dado por:

$$H_0: \beta_j = 0 \quad j = 0, 1, \dots, p$$

$$H_1: \beta_j \neq 0$$

$$(X'X)^{-1} = \begin{pmatrix} \frac{1}{12} & 0 & 0 & 0 \\ 0 & \frac{1}{12} & 0 & 0 \\ 0 & 0 & \frac{1}{12} & 0 \\ 0 & 0 & 0 & \frac{1}{12} \end{pmatrix}$$

 $C_{11} = C_{22} = C_{33} = \frac{1}{12} = 0,0833$

$$SQ_E = \mathbf{Y}'\mathbf{Y} - \hat{\boldsymbol{\beta}}' X'\mathbf{Y} = 9778,56 - 9729,84 = 48,72.$$

$$\hat{\sigma}^2 = \frac{SQ_E}{n - k - 1} = \frac{48,72}{12 - 3 - 1} = 6,09$$



Para β_1 : $t_0 = \frac{8,1}{\sqrt{6,09 * 0,0833}} = \frac{8,1}{\sqrt{0,507297}} = \frac{8,1}{0,7122479} = 11,37$

Para β_2 : $t_0 = \frac{-3,9}{\sqrt{6,09 * 0,0833}} = \frac{-3,9}{\sqrt{0,507297}} = \frac{-3,9}{0,7122479} = -5,47$

Para β_3 : $t_0 = \frac{1}{\sqrt{6,09 * 0,0833}} = \frac{1}{\sqrt{0,507297}} = \frac{1}{0,7122479} = 1,4$



Estatística do Teste

Rejeitamos H_0 se p-valor for menor que α , caso contrário não rejeitamos H_0 .

P-valor de AB = 0,198 > α , então a interação não é significativa.

P-valor de A = 0,000 < α , então A é significativo.

P-valor de B = 0,000 < α , então B é significativo.

Preditor	Efeitos	Coeficientes			
		Estimativa	Desvio Padrão	Estat. T	P-valor
Intercepto		27	0,712	37,90	2,58E-10
A		16,2	8,1	2,00	3,23E-06
B		-7,8	-3,9	2,00	0,00059
A:B		2	1	2,00	0,198

Análise dos Resíduos - Regressão

O resíduo (e_i) é dado pela diferença entre a variável resposta (Y_i) observada e a variável resposta estimada (\hat{Y}_i)

$$e_i = Y_i - \hat{Y}_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{1i} - \cdots - \hat{\beta}_p x_{pi}, \\ i = 1, \dots, n$$

Algumas técnicas gráficas para análise dos resíduos

Para a análise formal dos resíduos, podemos realizar os seguintes testes

Suposições para $Y = X\beta + \varepsilon$:

- O modelo é linear
- O erro tem média zero
- O erro tem variância σ^2 , constante
- Os erros são não correlacionados, ou seja, $Cov(\varepsilon_i, \varepsilon_j) = 0, (i \neq j)$
- Os erros $\varepsilon_i \sim N(0, \sigma^2)$

- Gráfico dos resíduos versus valores ajustados: verifica a homoscedasticidade do modelo, isto é, σ^2 constante.
- Gráfico dos resíduos versus a ordem de coleta dos dados: avaliar a hipótese de independência dos dados.
- Papel de probabilidade normal: verificar a normalidade dos dados.
- Gráfico dos Resíduos Studentizados versus valores ajustados: verifica se existem outliers em Y.
- Gráfico dos Resíduos Padronizados versus valores ajustados: verifica se existem outliers em Y.

- Teste de Normalidade em que detalhes estão contidos no conteúdo de Inferência.
- Teste de Durbin-Watson para testar independência dos resíduos.
- Teste de Breusch-Pagan e Goldfeld-Quandt para testar se os resíduos são homoscedásticos.
- Teste de falta de ajuste para verificar se o modelo ajustado é realmente linear.

Teste de Normalidade

Assumir a normalidade dos dados é o primeiro passo que tomamos para simplificar a análise. Para dar suporte a esta suposição, consideramos os testes:

-  Teste de Anderson-Darling;
-  Teste Kolmogorov-Smirnov;
-  Teste Shapiro-Wilk.
-  E o gráfico QQ-Plot

QQ-Plot e Teste de Anderson-Darling

Avaliamos a normalidade dos resíduos através do gráfico “papel de probabilidade” e do teste de Anderson-Darling. Consideramos a seguinte hipótese:

H_0 : os resíduos seguem uma distribuição Normal

H_1 : os resíduos Não seguem uma distribuição Normal

Diagnóstico de Outliers

Outlier é uma observação extrema, ou seja, é um ponto com comportamento diferente dos demais. Além de diagnosticar heteroscedasticidade, o gráfico de resíduos versus valores ajustados também auxilia na detecção de pontos atípicos.

A detecção de pontos atípicos tem por finalidade identificar:

- Outliers com relação a X
- Outliers com relação a Y
- Observações influentes

A seguir apresentamos algumas técnicas utilizadas:

Outliers em Y

- Resíduos Padronizados
- Resíduos Studentizados

Outliers em X

- Matriz chapéu H
- h_{ii} é o i-ésimo elemento da diagonal principal da matriz H e também é chamado de leverage da observação i.

Pontos Influentes

- DFFITS
- DFBETA
- Distância de Cook

Outliers em Y

Resíduos Padronizados



O resíduo padronizado, d_i , corresponde ao resíduo bruto dividido pelo erro padrão estimado dos resíduos, \sqrt{QME}

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

Se os erros têm distribuição normal, então aproximadamente 95% dos resíduos padronizados (d_i), devem estar no intervalo de $(-2,2)$. Resíduos fora desse intervalo podem indicar a presença de outliers.

Resíduos Studentizados



Vetor de resíduos

$$\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}} = \mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{Y} - \mathbf{H}\mathbf{Y} = (\mathbf{I} - \mathbf{H})\mathbf{Y},$$

No qual $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$



A matriz de covariâncias dos resíduos é,

$$Cov[\mathbf{e}] = \sigma^2(\mathbf{I} - \mathbf{H})$$



Resíduos Studentizados

$$r_i = \frac{e_i}{\sqrt{QME(I - h_{ii})}}, \quad i = 1, 2, \dots, n \text{ e } \hat{\sigma}^2 = QME$$

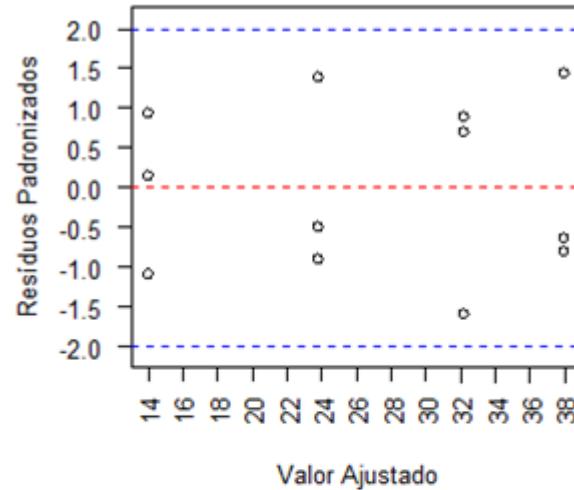
Exemplo

Exemplo 3.1.1: Estudar o efeito no tempo de uma determinada reação química com a variação de temperatura e concentração de um reagente.

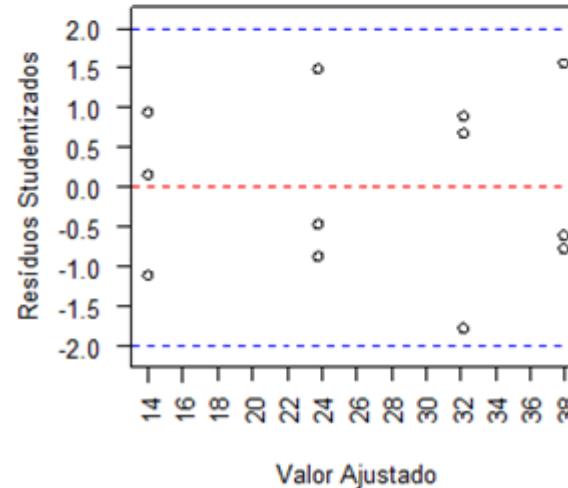
Outlier em Y

Resíduos	Resíduos Studentizados	Resíduos Padronizados
2,8	1,492	1,390
2,9	1,564	1,439
-2,2	-1,107	-1,092
1,8	0,881	0,893
-1,8	-0,881	-0,893
-1,6	-0,774	-0,794
1,9	0,936	0,943
-3,2	-1,795	-1,588
-1	-0,472	-0,496
-1,3	-0,620	-0,645
0,3	0,139	0,149
1,4	0,670	0,695

Res. Padronizado vs Valores Ajustados



Res. Studentizado vs Valores Ajustados



Teste de Outliers

Observação	t-Valor	p-valor	p-valor_Bonferroni
8	-1,7952776	0,11568113	1,388173525

Diagnóstico de Homoscedasticidade

Homoscedasticidade é o termo para designar variância constante dos erros ε_i para observações diferentes.

Caso a suposição de homoscedasticidade não seja válida, podemos listar alguns efeitos no ajuste do modelo:

- Os erros padrões dos estimadores, obtidos pelo Método dos Mínimos Quadrados, são incorretos e portanto a inferência estatística não é válida.
- Não podemos mais dizer que os Estimadores de Mínimos Quadrados são os melhores estimadores de mínima variância para , embora ainda possam ser não viciados.

A seguir apresentamos algumas técnicas utilizadas:

Teste de Goldfeld-Quandt



- é utilizado para testar a homoscedasticidade dos resíduos. Entre as limitações deste teste está a exigência de que a amostra seja relativamente grande.
- O teste consiste inicialmente em ordenar as observações de acordo com a variável explicativa que se acredita a responsável pela heteroscedasticidade.
- Após isso, divide-se a amostra ordenada em 3 partes de tal forma que a parte do meio tenha aproximadamente 20% dos dados e que as partes 1 e 3 tenham quantidade de dados semelhantes.
- Então, ajusta-se um modelo de regressão com os dados da parte 1 e outro modelo de regressão com os dados da parte 3
- testa-se a hipótese de que as variâncias dos erros em ambas regressões são iguais contra a hipótese de que a variância dos erros na parte 3 é maior do que a variância dos erros na parte 1, utilizando o teste F.

A estatística de teste neste caso é dada por

$$F_{GQ} = \frac{SQE^b / (n_3 - (p+1))}{SQE^a / (n_1 - (p+1))}$$

Técnica gráfica



Resíduos vs. Valores Ajustados

Diagnóstico de Independência

Para verificar se os resíduos são independentes, podemos utilizar técnicas gráficas e o testes de Durbin-Watson.

A seguir apresentamos algumas técnicas utilizadas:

- ❑ é utilizado para detectar a presença de auto correlação (dependência) nos resíduos de uma análise de regressão.
- ❑ teste é baseado na suposição de que os erros no modelo de regressão são gerados por um processo autoregressivo de primeira ordem, de acordo com

$$\varepsilon_i = \rho \varepsilon_{i-1} + a_i$$

em que ε é o termo do erro do modelo na i -ésima observação, $a_i \sim iid N(0, \sigma_a^2)$ e ρ ($|\rho| < 1$) é o parâmetro de autocorrelação.

- ❑ Testamos a presença de autocorrelação por meio das hipóteses

$$H_0: \rho = 0$$

$$H_1: \rho \neq 0$$

- ❑ Sendo o resíduo associado à i -ésima observação, temos que a estatística do teste de Durbin-Watson é dada por

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

em que $0 \leq dw \leq 4$

Teste de
Durbin-
Watson



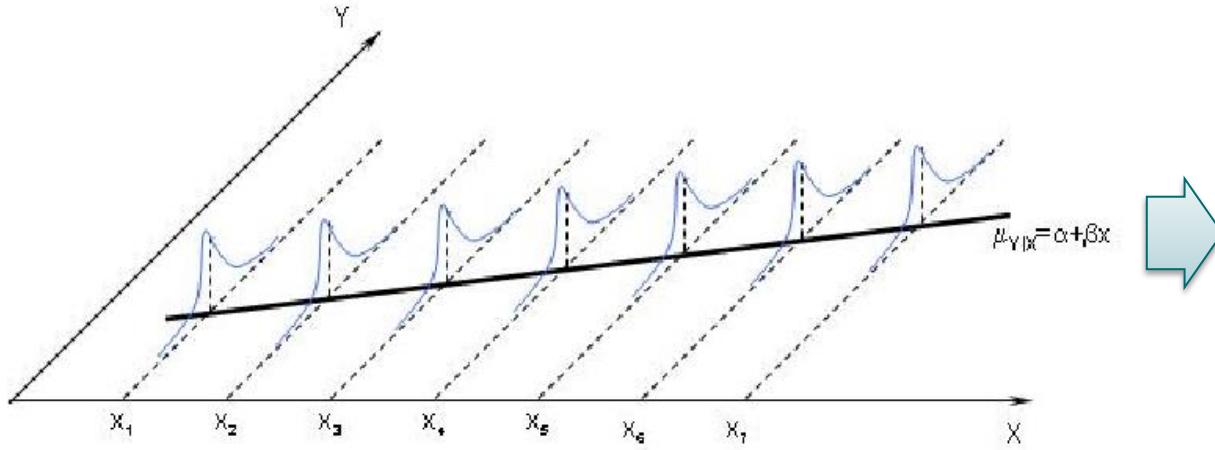
Técnica
gráfica



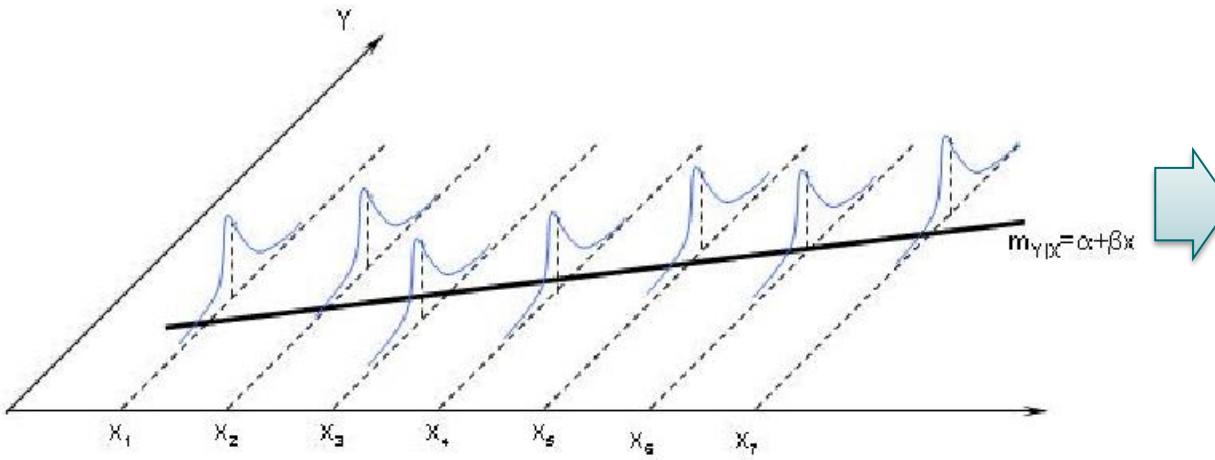
Resíduos vs. Ordem de Coleta

Teste da Falta de Ajuste (Lack of Fit)

Após o ajuste, é importante verificar se o modelo linear é adequado.



Reta de regressão
perfeitamente ajustada
sem Falta de Ajuste



Reta de regressão com
Falta de Ajuste

Teste da Falta de Ajuste (Lack of Fit)

A seguir, vamos fazer uma análise da falta de ajuste nos modelos simples e múltiplo. No caso simples, nosso modelo é:

$$y_{ij} = \beta_0 + \beta_1 x_i + \varepsilon \quad \begin{matrix} i = 1, 2, \dots, m \\ j = 1, 2, \dots, n_i \end{matrix}$$

Em que

- n_i : representa o número de observações para o i -ésimo valor de x .

Supondo que temos m diferentes valores da variável explicativa (x_1, x_2, \dots, x_m) e que temos n_i réplicas da variável resposta para cada valor da variável explicativa, ou seja,

$$\left\{ \begin{array}{l} x_1 \implies y_{11} \ y_{12} \ y_{13} \ \dots \ y_{1n_1} \\ x_2 \implies y_{21} \ y_{22} \ y_{23} \ \dots \ y_{1n_2} \\ x_3 \implies y_{31} \ y_{32} \ y_{33} \ \dots \ y_{1n_3} \\ \vdots \ \vdots \ \vdots \ \vdots \ \vdots \ \ddots \ \vdots \\ x_m \implies y_{m1} \ y_{m2} \ y_{m3} \ \dots \ y_{mn_m} \end{array} \right.$$

Para entendermos a quebra, tomamos,

$$e_{ij} = y_{ij} - \hat{y}_i = (y_{ij} - \bar{y}_{i.}) + (\bar{y}_{i.} - \hat{y}_i),$$

SQ_{PE}
(Erro Puro)

SQ_{LOF}
(Falta de Ajuste)

Análise da Falta de Ajuste nos modelos de Regressão Linear Simples

Teste de Hipótese

$$H_0: E(y_i) = \beta_0 + \beta_1 x_i \quad (\text{modelo linear adequado})$$
$$H_1: E(y_i) \neq \beta_0 + \beta_1 x_i \quad (\text{modelo linear inadequado})$$

Estatística do Teste

$$F_0 = \frac{\frac{SQ_{LOF}}{(m-2)}}{\frac{QM_{PE}}{(n-m)}} = \frac{QM_{LOF}}{QM_{PE}} \sim F_{(m-2, n-m)}$$

O valor de F_0 deve ser comparado com o quantil distribuição **F de Snedecor** com $(m - 2)$ graus de liberdade no numerador e $(n - m)$ graus de liberdade no denominador para um determinado nível de significância α

Fonte	GL	Soma de quadrados	Quadrado médio	F_0
Regressão	1	$SQR = \hat{\beta}_1 S_{xy}$	$QMR = \frac{SQR}{SQE}$	$F_0 = \frac{QMR}{QME}$
Resíduo	$n - 2$	$SQE = \sum_{i=1}^m \sum_{j=1}^{n_i} (y_{ij} - \hat{y}_i)^2$	$QME = \frac{SQE}{(n-2)}$	
Falta de ajuste	$m - 2$	$SQ_{LOF} = \sum_{i=1}^m n_i (\hat{y}_i - \bar{y}_{i.})^2$	$QM_{LOF} = \frac{SQ_{LOF}}{(m-2)}$	$F_0 = \frac{QM_{LOF}}{QM_{PE}}$
Erro puro	$n - m$	$SQ_{PE} = \sum_{i=1}^m \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i.})^2$	$QM_{PE} = \frac{SQ_{PE}}{(n-m)}$	
Total	$n - 1$	$SQT = S_{yy}$		

Exemplo

Exemplo 3.1.1: Estudar o efeito no tempo de uma determinada reação química com a variação de temperatura e concentração de um reagente.

Utilizando o modelo com as interações, temos:

$$SQE = \sum_{i=1}^m \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i.})^2 + \sum_{i=1}^m \sum_{j=1}^{n_i} (\bar{y}_{i.} - \hat{y}_i)^2 = 48,72$$

$$QM_{PE} = \frac{SQ_{PE}}{n - m} = \frac{48,72}{8} = 6,09$$

$$QM_{LOF} = \frac{SQ_{LOF}}{m - 2} = \frac{-3,55E^{-14}}{0}$$

Fonte	SQ	GL	QM	Estatística	P-valor
Regressão	969,84	2	484,92	79,63	
Resíduo	48,72	8	6,09		
Falta de Ajuste	-3,55E-14	0			
Erro Puro	48,72	8	6,09		
Total	1091,28	18			

Notamos neste exemplo que se pegarmos o modelo cheio, ou seja, com as interações, o teste de falta de ajuste não é calculado.

Exemplo

Exemplo 3.1.1: Estudar o efeito no tempo de uma determinada reação química com a variação de temperatura e concentração de um reagente.

Utilizando o modelo sem as interações, temos:

$$SQE = \sum_{i=1}^m \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i.})^2 + \sum_{i=1}^m (\bar{y}_{i.} - \hat{y}_i)^2 = 60,72$$

$$QM_{PE} = \frac{SQ_{PE}}{n - m} = \frac{48,72}{8} = 6,09$$

$$QM_{LOF} = \frac{SQ_{LOF}}{m - 2} = \frac{12}{1} = 12$$

$$F_0 = \frac{QM_{LOF}}{QM_{PE}} = \frac{12}{6,9} = 1,97 \sim F_{(1;8)}$$

$$P\text{-valor} = P[F_{(1;8)} > F_0] = 0,2$$

Fonte	SQ	GL	QM	Estatística	P-valor
Regressão	969,84	2	484,92	79,63	
Resíduo	60,72	9	6,75		
Falta de Ajuste	12	1	12	1,97	0,2
Erro Puro	48,72	8	6,09		
Total	1091,28	20			

Análise da Falta de Ajuste nos modelos de Regressão Linear Múltipla

Teste de Hipótese

$$H_0: E(y_i) = \beta_0 + \beta_1 x_i \quad (\text{modelo linear adequado})$$
$$H_1: E(y_i) \neq \beta_0 + \beta_1 x_i \quad (\text{modelo linear inadequado})$$

Estatística do Teste

$$F_0 = \frac{\frac{SQ_{LOF}}{(m-p-1)}}{\frac{QM_{PE}}{(n-m)}} = \frac{QM_{LOF}}{QM_{PE}} \sim F_{(m-p-1, n-m)}$$

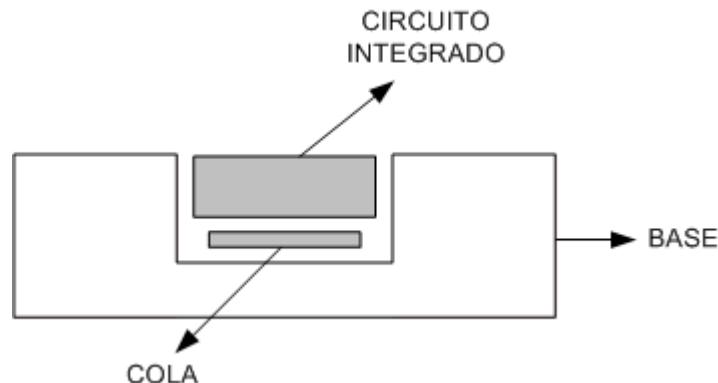
O valor de F_0 deve ser comparado com o quantil distribuição **F de Snedecor** com $(m - p - 1)$ graus de liberdade no numerador e $(n - m)$ graus de liberdade no denominador para um determinado nível de significância α

Fonte	GL	Soma de quadrados	Quadrado médio	F_0
Regressão	1	$SQR = \hat{\beta}_1 S_{xy}$	$QMR = \frac{SQR}{SQE}$	$F_0 = \frac{QMR}{QME}$
Resíduo	$n - 2$	$SQE = \sum_{i=1}^m \sum_{j=1}^{n_i} (y_{ij} - \hat{y}_i)^2$	$QME = \frac{SQE}{(n-2)}$	
Falta de ajuste	$m - 2$	$SQ_{LOF} = \sum_{i=1}^m n_i (\hat{y}_i - \bar{y}_{i.})^2$	$QM_{LOF} = \frac{SQ_{LOF}}{(m-2)}$	$F_0 = \frac{QM_{LOF}}{QM_{PE}}$
Erro puro	$n - m$	$SQ_{PE} = \sum_{i=1}^m \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i.})^2$	$QM_{PE} = \frac{SQ_{PE}}{(n-m)}$	
Total	$n - 1$	$SQT = S_{yy}$		

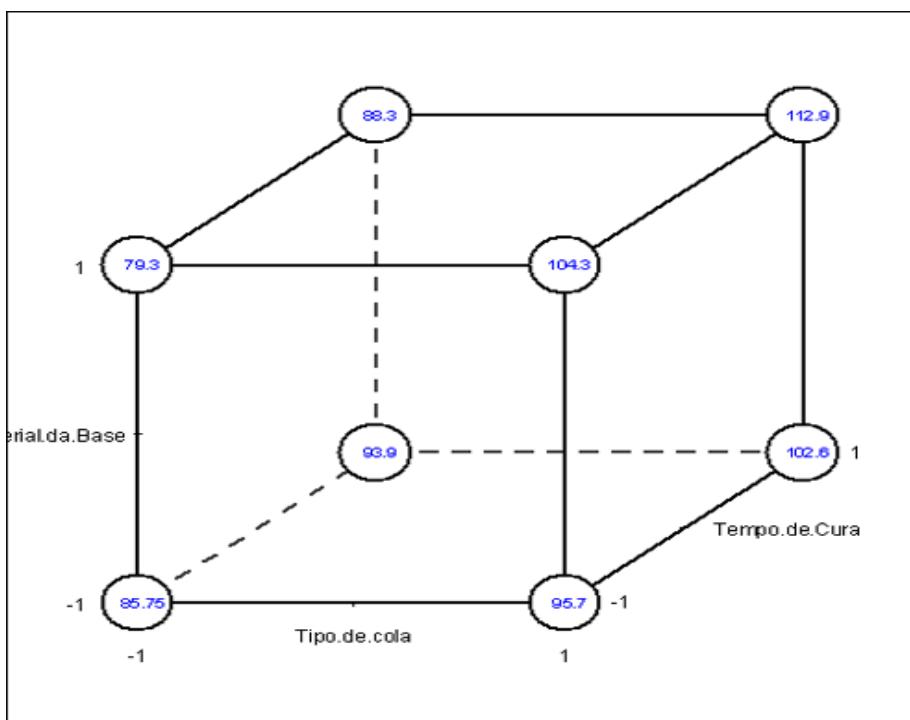
Experimentos Fatoriais Completos 2^3

Exemplo

Exemplo 3.2.1: Processo de colagem de um circuito integrado em uma base.



Variável Resposta	Y : Resistência da cola (libras)
Fatores	A: Tipo de cola (níveis A_{-1} : marca atual e A_{+1} : marca nova) B: Material da base (níveis B_{-1} : atual e B_{+1} : nova especificação) C: Tempo de cura a $100^{\circ}C$ (níveis C_{-1} : 90 min e C_{+1} : 120 min)
Número de tratamentos:	$2^3 = 8$
Número de réplicas:	2
Número de corridas:	$8 * 2 = 16$



Exemplo

Gráfico de Efeitos Principais

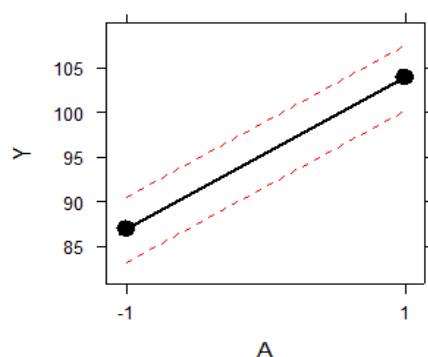


Gráfico de Efeitos Principais

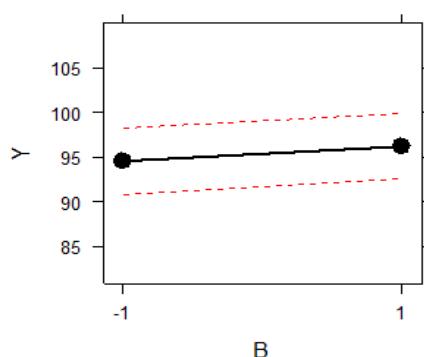


Gráfico de Interações

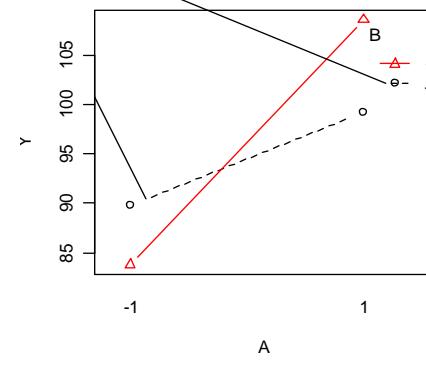


Gráfico de Interações

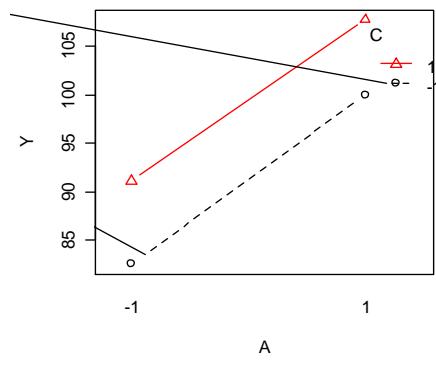


Gráfico de Efeitos Principais

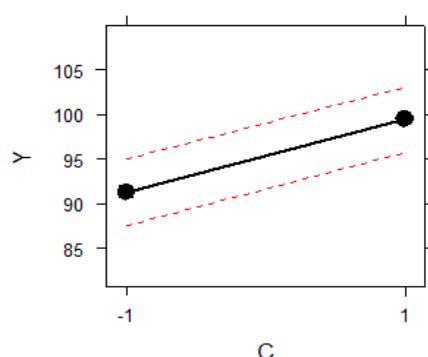
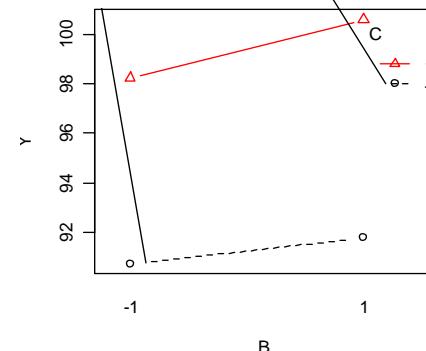


Gráfico de Interações



A	Limite Inferior	Efeito	Limite Superior
-1	83,14	86,81	90,48
1	100,20	103,88	107,55
B	Limite Inferior	Efeito	Limite Superior
-1	90,82	94,49	98,16
1	92,53	96,20	99,87
C	Limite Inferior	Efeito	Limite Superior
-1	87,59	91,26	94,93
1	95,75	99,43	103,10

Exemplo

Na forma matricial temos

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

$$X = \begin{pmatrix} (A) & (B) & (C) & (AB) & (AC) & (BC) & (ABC) \\ 1 & -1 & -1 & -1 & +1 & +1 & +1 & -1 \\ 1 & +1 & -1 & -1 & -1 & -1 & +1 & +1 \\ 1 & -1 & +1 & -1 & -1 & +1 & -1 & +1 \\ 1 & +1 & +1 & -1 & +1 & -1 & -1 & -1 \\ 1 & -1 & -1 & +1 & +1 & -1 & -1 & +1 \\ 1 & +1 & -1 & +1 & -1 & +1 & -1 & -1 \\ 1 & -1 & +1 & +1 & -1 & -1 & +1 & -1 \\ 1 & +1 & +1 & +1 & +1 & +1 & +1 & +1 \\ 1 & -1 & -1 & -1 & +1 & +1 & +1 & -1 \\ 1 & +1 & -1 & -1 & -1 & +1 & +1 & +1 \\ 1 & -1 & +1 & -1 & -1 & -1 & +1 & +1 \\ 1 & +1 & +1 & -1 & +1 & -1 & -1 & -1 \\ 1 & -1 & -1 & +1 & -1 & -1 & +1 & +1 \\ 1 & +1 & -1 & +1 & +1 & -1 & -1 & -1 \\ 1 & -1 & +1 & +1 & -1 & +1 & -1 & +1 \\ 1 & +1 & +1 & +1 & +1 & +1 & +1 & +1 \end{pmatrix}, Y = \begin{pmatrix} 82,90 \\ 95,00 \\ 80,00 \\ 105,00 \\ 93,10 \\ 104,00 \\ 89,70 \\ 114,00 \\ 88,60 \\ 96,40 \\ 78,60 \\ 103,60 \\ 94,70 \\ 101,20 \\ 86,90 \\ 111,80 \end{pmatrix}$$

Tratamento	A	B	C	AB	AC	BC	ABC	Y_1	Y_2	\bar{Y}
0	-1	-1	-1	1	1	1	-1	82,9	88,6	85,75
a	1	-1	-1	-1	-1	1	1	95	96,4	95,70
b	-1	1	-1	-1	1	-1	1	80	78,6	79,30
ab	1	1	-1	1	-1	-1	-1	105	103,6	104,30
c	-1	-1	1	1	-1	-1	1	93,1	94,7	93,90
ac	1	-1	1	-1	1	-1	-1	104	101,2	102,60
bc	-1	1	1	-1	-1	1	-1	89,7	86,9	88,30
abc	1	1	1	1	1	1	1	114	111,8	112,90

$$\hat{\boldsymbol{\beta}} = (X'X)^{-1}X'Y = \begin{pmatrix} 95,3438 \\ 8,5312 \\ 0,8562 \\ 4,0812 \\ 3,8687 \\ -0,2062 \\ 0,3188 \\ 0,1062 \end{pmatrix}$$



Modelo Ajustado

$$Y = 95,34 + 8,53X_1 + 0,85X_2 + 4,08X_3 + 3,86X_4 - 0,206X_5 + 0,318X_6 + 0,106X_7$$

Exemplo

Exemplo 3.2.1: Processo de colagem de um circuito integrado em uma base.

Coeficientes					
Predictor	Efeitos	Estimativa	Desvio Padrão	Estat. T	P-valor
Intercepto		95,344	0,490	194,604	5,44E-16
A	17,063	8,531	0,490	17,413	1,21E-07
B	1,7125	0,856	0,490	1,748	0,12
C	8,1625	4,081	0,490	8,330	3,26E-05
A:B	7,7375	3,869	0,490	7,896	4,80E-05
A:C	-0,4125	-0,206	0,490	-0,421	0,68
B:C	0,6375	0,319	0,490	0,651	0,53
A:B:C	0,2125	0,106	0,490	0,217	0,83

Tabela da Anova						
Fatores	G.L.	Soma de Quadrados	Quadrado Médio	Estat. F	P-valor	
A	1	1164,52	1164,52	303,21	1,21E-07	
B	1	11,73	11,73	3,05	0,12	
C	1	266,51	266,51	69,39	3,26E-05	
A:B	1	239,48	239,48	62,35	4,80E-05	
A:C	1	0,68	0,68	0,18	0,68	
B:C	1	1,63	1,63	0,42	0,53	
A:B:C	1	0,18	0,18	0,047	0,83	
Residuals	8	30,72	3,84			

Relembrando: A é o tipo de cola, B é o material da base, C é o tempo de cura.

Como os p-valores para A, C e AB são menores que $\alpha = 5\%$ esses fatores são significativos, porém o fator B não é significativo, como notou-se nos gráficos de efeitos principais.

Conclusões:

- Os fatores A, C e AB são significativos, mas B não;
- O coeficiente de regressão de A é maior, em módulo, que o de AB, então podemos olhar apenas para os níveis de A e C;
- Como os coeficientes da regressão de ambos são positivos, concluímos que A e C são melhores no nível +1
- Como B não é significativo, é melhor conservar o nível atual -1.
- Portanto a melhor configuração seria A+;B-,C+.

Experimentos Fatoriais sem Réplicas

Experimentos Fatoriais sem Réplicas

Introdução



Em algumas situações o número de combinações dos fatores do experimento é grande. Nestes casos, recursos podem estar disponíveis apenas para uma única execução do projeto, ou seja, o experimento não possuirá réplicas.



Um risco evidente quando se conduz um experimento sem réplicas é que o modelo ajustado pode levar a conclusões erradas. Além disto, neste caso não há estimativa interna de erro (erro puro).

Existem alguns métodos para tratarem estes experimento, dentre eles citamos os métodos de Daniel e Lenth que são métodos objetivos para decidir quais efeitos são significativos na análise de experimentos sem réplicas, nas situações em que o modelo está saturado e assim, não há graus de liberdade para estimar a variância do erro.

Experimentos Fatoriais sem Réplicas

Gráfico da Probabilidade Normal e Half-Normal

No estudo de experimento fatoriais sem réplicas Cuthbert Daniel (1959) propôs um método que avalia estes efeitos ativos.

A ideia de Daniel é bastante utilizada até os dias atuais por ser simples e conseguir apontar a direção correta dos efeitos em grande parte dos experimentos.

A aplicação eficaz desses gráficos depende do fato das estimativas dos efeitos terem a mesma variância, e os pontos em que temos "efeitos esparsos" são detectados pelo método.

Segundo Daniel, esperamos que apenas uma pequena fração dos contrastes sejam ativos dentre todos aqueles envolvidos no estudo. Nestes gráficos, os efeitos cujos pontos estiverem claramente afastados de uma reta imaginária, formada pela nuvem de pontos, serão julgados ativos.

Experimentos Fatoriais sem Réplicas

Sejam $\hat{c}_1, \hat{c}_2, \dots, \hat{c}_n$ os n efeitos estimados, onde $\hat{c}_{(i)}$, o i -ésimo dos n efeitos ordenados, $\hat{c}_{(1)}, \hat{c}_{(2)}, \dots, \hat{c}_{(n)}$.

Dessa forma, o gráfico de probabilidade normal pode ser obtido dispondo-se os pontos em um gráfico cujas coordenadas (x, y) são dadas por:

$$\left(\hat{c}_{(i)}, \Phi^{-1} \left[\frac{(i - 0,5)}{n} \right] \right)$$

Em que $\Phi^{-1}(.)$ é a função de distribuição acumulada da normal padrão.

Alguns autores preferem utilizar o gráfico half-normal cujas coordenadas são dadas por:

$$\left(|\hat{c}_{(i)}|, \Phi^{-1} \left[0,5 + \frac{(i - 0,5)}{n} \right] \right)$$

Uma das vantagens de utilizar o half-normal é o fato de que os efeitos possivelmente ativos vão se apresentar no canto superior direito do gráfico.

Exemplo

Exemplo 3.3.1 Um determinado produto químico é produzido em um vaso de pressão. Com o objetivo de estudar quais fatores influenciam na taxa de filtração do produto (Y), foi realizado um experimento fatorial em que se considerou 4 fatores: A (temperatura), B (pressão), C (concentração de formaldeído) e D (velocidade de agitação). Cada fator é observado em dois níveis.

Tratamento	A	B	C	D	Y
0	-1	-1	-1	-1	45
a	1	-1	-1	-1	71
b	-1	1	-1	-1	48
ab	1	1	-1	-1	65
c	-1	-1	1	-1	68
ac	1	-1	1	-1	60
bc	-1	1	1	-1	80
abc	1	1	1	-1	65
d	-1	-1	-1	1	43
ad	1	-1	-1	1	100
bd	-1	1	-1	1	45
abd	1	1	-1	1	104
cd	-1	-1	1	1	75
acd	1	-1	1	1	86
bcd	-1	1	1	1	70
abcd	1	1	1	1	96

Exemplo

O efeito é obtido da seguinte forma:

$$c = \bar{y}_+ - \bar{y}_-$$

Efeitos	Estimativa
Intercepto	70,0625
A	21,625
B	3,125
C	9,875
D	14,625
A:B	0,125
A:C	-18,13
B:C	2,375
A:D	16,625
B:D	-0,375
C:D	-1,125
A:B:C	1,875
A:B:D	4,125
A:C:D	-1,625
B:C:D	-2,625
A:B:C:D	1,375
	0,6875

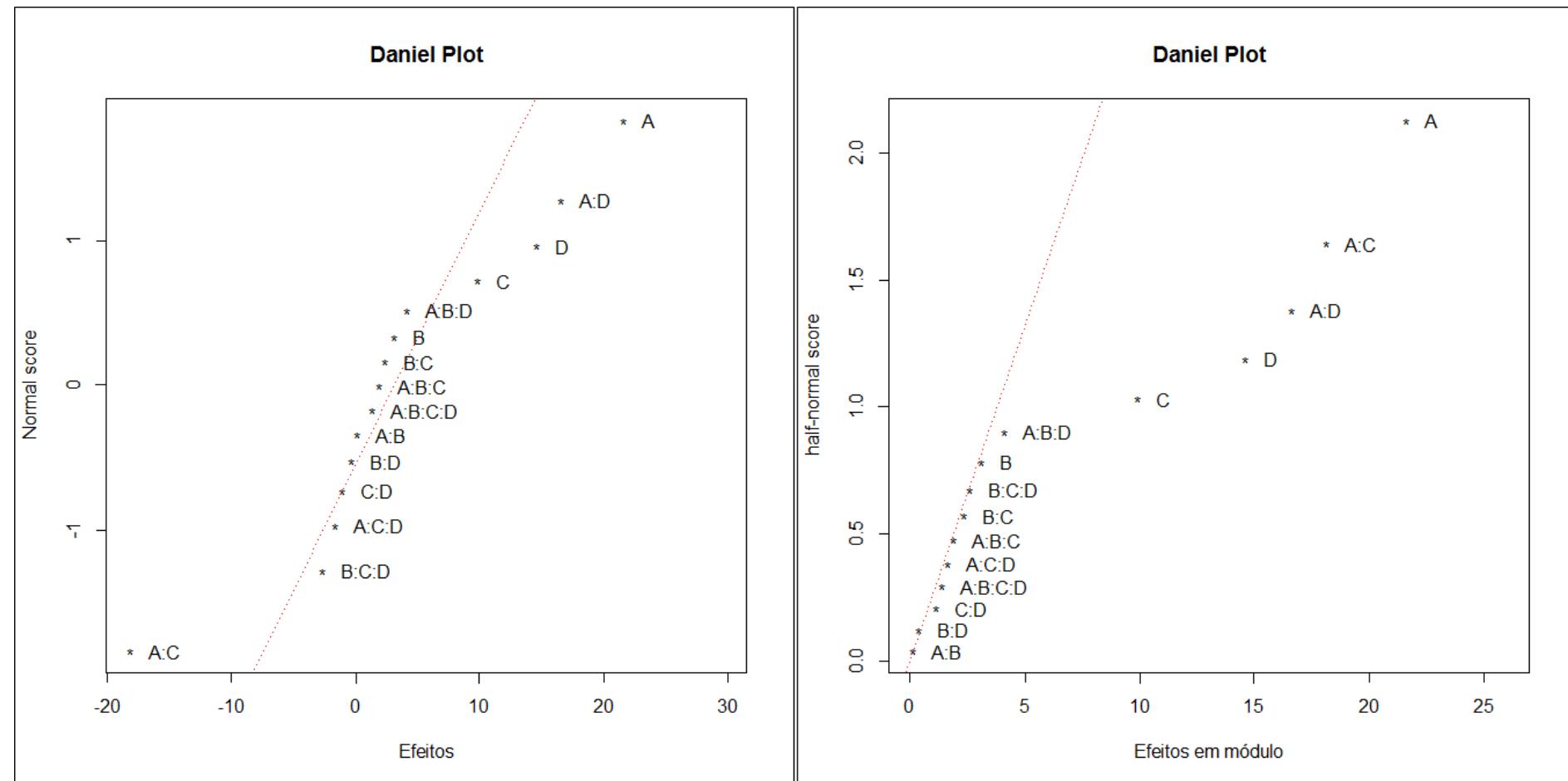
	Médias			
	A	B	C	D
-1	59,25	68,5	65,12	62,75
1	80,87	71,62	75	77,37

$$c = \bar{y}_+ - \bar{y}_- = 80,87 - 59,25 = 21,625$$

$$c = \bar{y}_+ - \bar{y}_- = 71,62 - 68,5 = 3,125$$

⋮

Exemplo



Experimentos Fatoriais sem Réplicas

Método de Lenth



O método de Lenth tem sido considerado como um método muito eficiente, quando trabalhamos com análise de experimentos fatoriais sem réplicas.



Um ponto para análise destes experimentos, é o estudo de um número grande de contrastes e que as estimativas destes contrastes tenham a mesma variabilidade.



O método de Lenth, assim como Daniel, parte do princípio de que tenhamos apenas poucos "efeitos esparsos" (efeitos dispersos), que o autor trata como efeitos ativos (diferente de zero), ou seja, efeitos significativamente não nulos do ponto de vista estatístico.

Experimentos Fatoriais sem Réplicas

Considere um factorial com dois níveis e suponha k_1, k_2, \dots, k_m contrastes e c_1, c_2, \dots, c_m efeitos estimados independentes com mesma variância $\sigma^2 \sim N(k_i, \sigma^2)$. Sendo N o número de observações. Desta forma temos:

$$c = \bar{y}_+ - \bar{y}_-$$

Em que:

\bar{y}_+ é a média das $\frac{N}{2}$ observações no nível alto

\bar{y}_- é a média das $\frac{N}{2}$ observações no nível baixo

Calculamos o pseudo erro padrão (PSE) como sendo:

$$PSE = 1,5 * mediana\{|c_j| : |c_j| \leq 2,5s_0\}, j = 1, \dots, m$$

Em que:

$$s_0 = 1,5 * mediana\{|c_j|\}$$

Notamos que s_0 e PSE são bastante similares, com uma pequena diferença na mediana do PSE que é mais restrita. Esta restrição é devido aos pontos ativos e é descrita no artigo R, que é feita para obtermos estimativas consistentes para σ .

Experimentos Fatoriais sem Réplicas

Em relação ao critério de decisão de quais efeitos são significativos, definimos uma margem de erro dos contrastes C_i denotada por ME. O valor da margem de erro é dada por:

$$ME = t_{\left\{1 - \frac{\alpha}{2}; d\right\}} * PSE$$

Em que:

$$d = \frac{m}{3}$$

$t_{\left\{1 - \frac{\alpha}{2}; d\right\}}$ é o quantil $\left(1 - \frac{\alpha}{2}\right)$ da t-student com d gl

α é o nível de significância adotado.

Quando há um grande número de contrastes m esperamos que uma ou duas estimativas de contrastes não significativos excedam o valor de ME, conduzindo a uma falsa conclusão. Desta forma, a fim de tratar estes casos, é definida uma margem de erro simultânea, que é dada por:

$$SME = t_{\gamma; d} * PSE$$

Em que: $\gamma = \frac{\left(1 + 0,95^{\frac{1}{m}}\right)}{2}$

Experimentos Fatoriais sem Réplicas

Critério para avaliar os efeitos

Intervalo de Confiança

$$LI = \hat{c}_j - ME$$

$$j = 1, \dots, k$$

$$LS = \hat{c}_j + ME$$

Teste de Hipótese

$$H_0: c_j = 0$$

$$H_1: c_j \neq 0, \text{ para algum } j = 1, \dots, k$$

Estatística de Lenth

$$T_{Lj} = \frac{|\hat{c}_j|}{PSE} \sim t_{(d)}$$

P-valor

$$p\text{-valor} = 2 * P[t_d > |T_{Lj}| | H_0]$$

Exemplo

Exemplo 3.3.1 Um determinado produto químico é produzido em um vaso de pressão. Com o objetivo de estudar quais fatores influenciam na taxa de filtração do produto (Y), foi realizado um experimento fatorial em que se considerou 4 fatores: A (temperatura), B (pressão), C (concentração de formaldeído) e D (velocidade de agitação). Cada fator é observado em dois níveis.

Tratamento	A	B	C	D	Y
0	-1	-1	-1	-1	45
a	1	-1	-1	-1	71
b	-1	1	-1	-1	48
ab	1	1	-1	-1	65
c	-1	-1	1	-1	68
ac	1	-1	1	-1	60
bc	-1	1	1	-1	80
abc	1	1	1	-1	65
d	-1	-1	-1	1	43
ad	1	-1	-1	1	100
bd	-1	1	-1	1	45
abd	1	1	-1	1	104
cd	-1	-1	1	1	75
acd	1	-1	1	1	86
bcd	-1	1	1	1	70
abcd	1	1	1	1	96

Exemplo

Com a estimativa dos efeitos, calculamos:

Efeitos	Estimativa
Intercepto	70,0625
A	21,625
B	3,125
C	9,875
D	14,625
A:B	0,125
A:C	-18,13
B:C	2,375
A:D	16,625
B:D	-0,375
C:D	-1,125
A:B:C	1,875
A:B:D	4,125
A:C:D	-1,625
B:C:D	-2,625
A:B:C:D	1,375

$$S_0 = 1,5 * \text{mediana}\{21,625; 125; \dots; 2,625; 1,375\} = 3,938$$

$$PSE = 1,5 * \text{mediana}\{|c_j| : |c_j| \leq 2,5 * 3,938\} = 2,625$$

$$ME = t_{\left\{1 - \frac{0,05}{2}; 5\right\}} * 2,625 = 2,571 * 2,625 = 6,748$$

$$\begin{aligned} SME &= t_{0,998;5} * 2,625 = 5,219 * 2,625 \\ &= 13,699 \end{aligned}$$

Em que:

$$\gamma = \frac{\left(1 + 0,95^{\frac{1}{15}}\right)}{2} = 0,998$$

Exemplo

	Efeitos	Estimativa
Intercepto		70,0625
A	21,625	10,8125
B	3,125	1,5625
C	9,875	4,9375
D	14,625	7,3125
A:B	0,125	0,0625
A:C	-18,13	-9,0625
B:C	2,375	1,1875
A:D	16,625	8,3125
B:D	-0,375	-0,1875
C:D	-1,125	-0,5625
A:B:C	1,875	0,9375
A:B:D	4,125	2,0625
A:C:D	-1,625	-0,8125
B:C:D	-2,625	-1,3125
A:B:C:D	1,375	0,6875

O efeito é obtido da seguinte forma:

$$c = \bar{y}_+ - \bar{y}_- = 80,87 - 59,25 = 21,625$$

Vamos adotar o efeito A para calcularmos o intervalo de confiança e a estatística t.

$$LI = \hat{c}_1 - ME = 21,625 - 6,748 = 14,877$$

$$LS = \hat{c}_1 + ME = 21,625 + 6,748 = 28,372$$

$$T_{Lj} = \frac{|\hat{c}_1|}{PSE} = \frac{21,625}{2,625} = 8,238$$

$$p\text{-valor} = 2 * P[t_d > |T_{Lj}| | H_0] = 0,0002$$

Exemplo

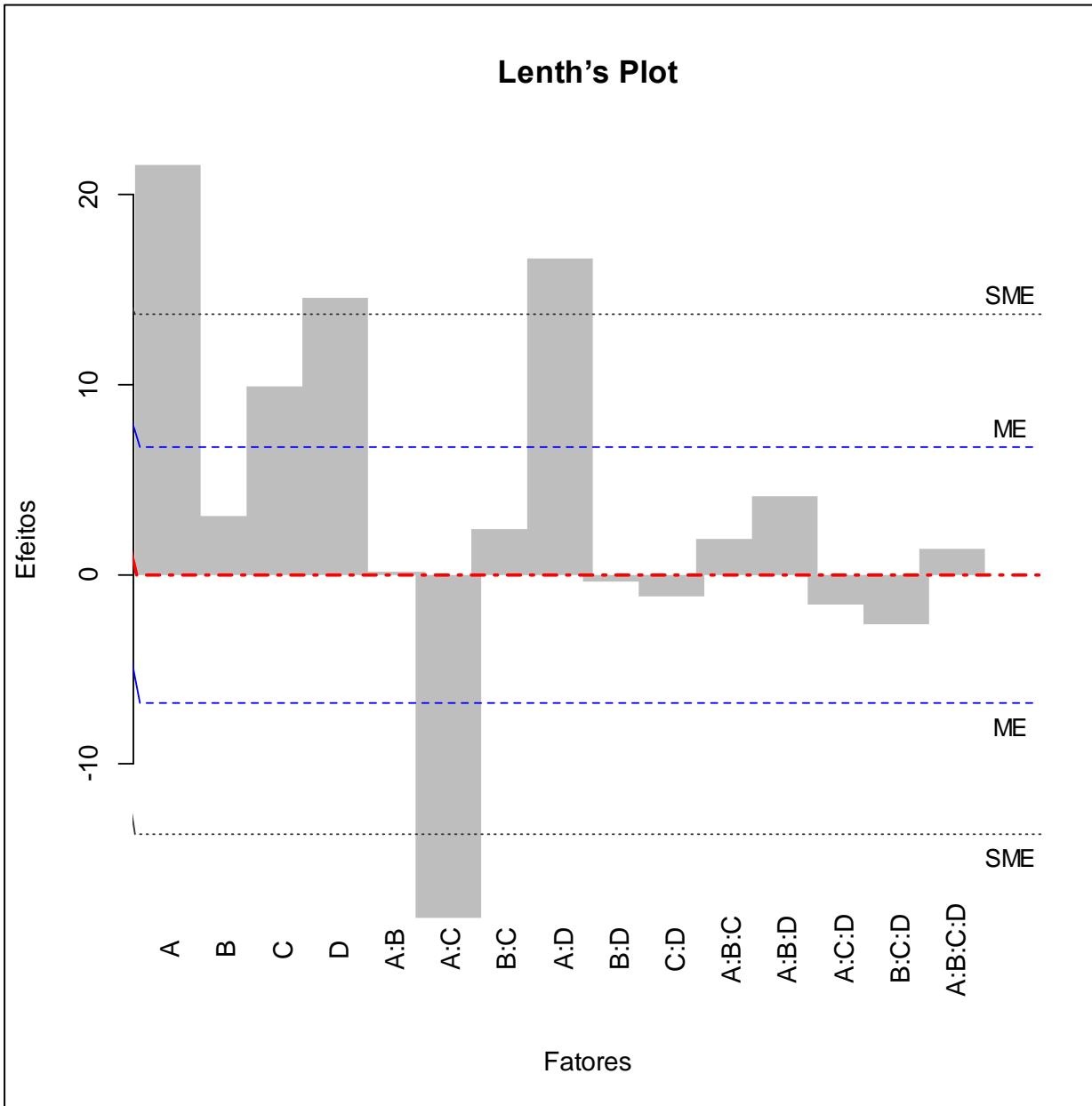
MODELO DE REGRESSÃO LINEAR

DADOS DO PROCESSO

Efeitos	Estimativa	Limite Inferior	Limite Superior	Estatística t	P-valor
Intercepto	70,0625				
A	21,625	10,8125	14,87722268	28,37277732	8,238095238 0,000214738
B	3,125	1,5625	-3,622777319	9,872777319	1,19047619 0,143646542
C	9,875	4,9375	3,127222681	16,62277732	3,761904762 0,006565881
D	14,625	7,3125	7,877222681	21,37277732	5,571428571 0,001282684
A:B	0,125	0,0625	-6,622777319	6,872777319	0,047619048 0,481931684
A:C	-18,13	-9,0625	-24,87277732	-11,37722268	6,904761905 0,000488167
B:C	2,375	1,1875	-4,372777319	9,122777319	0,904761905 0,203536033
A:D	16,625	8,3125	9,877222681	23,37277732	6,333333333 0,000723737
B:D	-0,375	-0,1875	-7,122777319	6,372777319	0,142857143 0,445990739
C:D	-1,125	-0,5625	-7,872777319	5,622777319	0,428571429 0,343035335
A:B:C	1,875	0,9375	-4,872777319	8,622777319	0,714285714 0,253500253
A:B:D	4,125	2,0625	-2,622777319	10,87277732	1,571428571 0,088442809
A:C:D	-1,625	-0,8125	-8,372777319	5,122777319	0,619047619 0,281498515
B:C:D	-2,625	-1,3125	-9,372777319	4,122777319	1 0,181608734
A:B:C:D	1,375	0,6875	-5,372777319	8,122777319	0,523809524 0,311395477

	Valor
alpha	0,05
PSE	2,625
ME	6,7478
SME	13,699
t.crit	2,5706

Exemplo



Experimentos Fatoriais em Blocos

Experimentos Fatoriais em Blocos

Introdução

- Conforme o número de fatores envolvidos em um experimento, pode tornar-se inviável realizar todas as corridas.
- Em outros casos, pode ser desejável varrer todas as condições do experimento para garantir um tratamento igual em todas as condições.
- A técnica de delineamento de experimentos utilizada neste caso é chamada de blocos.

Experimento Fatorial 2^k

- Para realizar um experimento fatorial 2^k com n réplicas sob condições não homogêneas podemos considerar cada conjunto dessas condições como um bloco e realizar as réplicas em blocos diferentes.
- Para situações em que não é possível fazer uma réplica completa de um experimento fatorial em um único bloco utiliza-se uma técnica denominada de confundimento.
- Esta técnica faz com que certos efeitos de interação fiquem indistinguíveis dos blocos, ou confundidos com os blocos.
- Considerando um planejamento fatorial 2^k em 2^p blocos, sendo $p < k$, pode-se delinear este experimento em dois blocos, quatro blocos, oito blocos e assim por diante.

Experimentos Fatoriais em Blocos

Dois Blocos



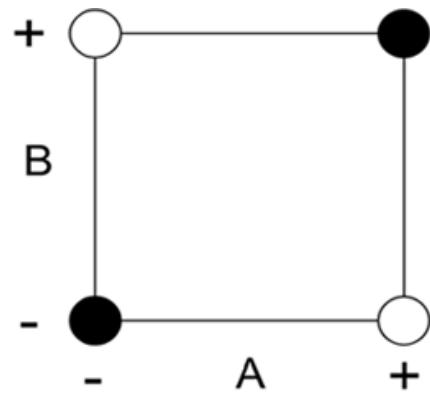
Supondo um experimento factorial 2^2 com uma única réplica tem-se quatro possíveis combinações.



Muitas vezes, os recursos disponíveis, como por exemplo, a matéria prima não é suficiente para realizar as quatro combinações, mas apenas duas delas.



Neste caso, pode-se considerar cada tipo de matéria prima como um bloco, o que permite alocar duas combinações para cada bloco.



No bloco 1 observa-se as combinações $(0, ab)$ e no bloco 2 as combinações $[a,b]$, sendo que a ordem das corridas é aleatória.



Experimentos Fatoriais em Blocos

Dois Blocos

Os procedimentos para estimar os efeitos principais de A e B são os mesmos usados para estimativas em experimentos que não são delineados em blocos.

$$A = \frac{1}{2} [ab + a - b - (0)]$$

$$B = \frac{1}{2} [ab + b - a - (0)]$$

$$AB = \frac{1}{2} [ab + (0) - a - b]$$

A prática usual determina que confundamos as interações de maior ordem com os blocos.

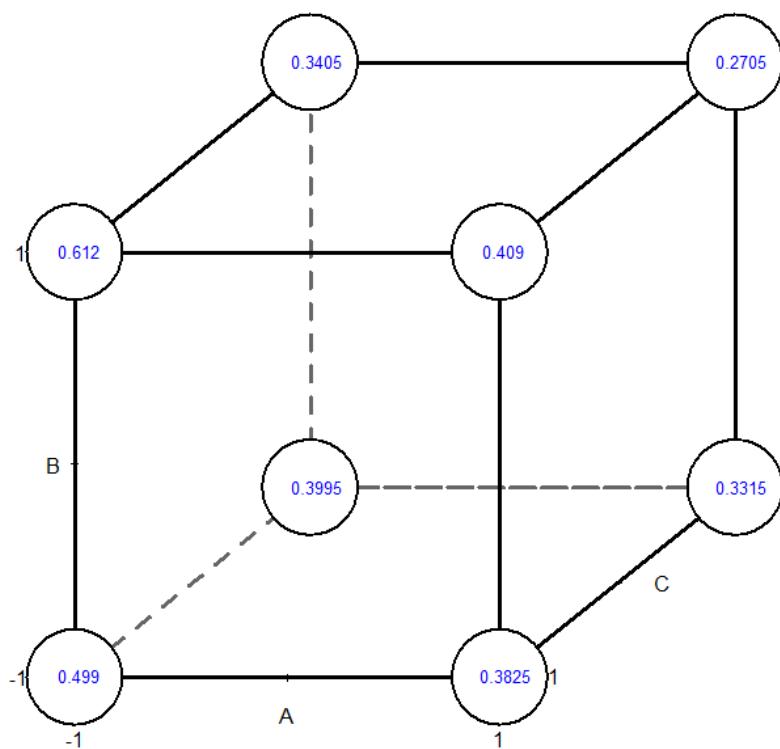
Tratamento	Efeitos fatoriais				
	I	A	B	AB	Bloco
(0)	+	-	-	+	1
a	+	+	-	-	2
b	+	-	+	-	2
ab	+	+	+	+	1

Exemplo

Exemplo 4.1.1.1: Uma empresa fabricante de bolos quer reduzir as reclamações relacionadas a dureza de sua massa. Para isso delineou um experimento 2^3 , com duas réplicas, com os fatores definidos da seguinte forma:

- A - Açúcar;
- B - Leite;
- C - Fermento.

A	B	C	Leveza
-1	-1	-1	0,499
1	-1	-1	0,3825
-1	-1	1	0,3995
1	-1	1	0,3315
-1	1	-1	0,612
1	1	-1	0,409
-1	1	1	0,3405
1	1	1	0,2705



Réplica	Bloco	A	B	C	Leveza
1	1	-1	-1	-1	0,539
1	1	1	1	-1	0,446
1	1	1	1	1	0,306
1	1	-1	1	1	0,371
1	2	1	-1	-1	0,411
1	2	-1	1	-1	0,602
1	2	-1	-1	1	0,369
1	2	1	1	1	0,295
2	1	-1	-1	-1	0,459
2	2	-1	-1	1	0,43
2	1	1	1	-1	0,372
2	2	1	1	1	0,246
2	2	1	-1	-1	0,354
2	1	1	-1	1	0,622
2	1	-1	1	1	0,357
2	1	-1	1	1	0,31

Exemplo

Exemplo 4.1.1.1: Uma empresa fabricante de bolos quer reduzir as reclamações relacionadas a dureza de sua massa. Para isso delineou um experimento 2^3 , com duas réplicas, com os fatores definidos da seguinte forma:

- A - Açúcar;
- B - Leite;
- C - Fermento.

Gráfico de Efeitos

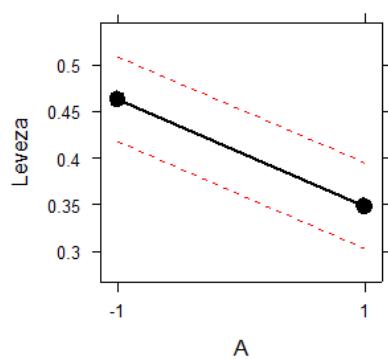


Gráfico de Efeitos

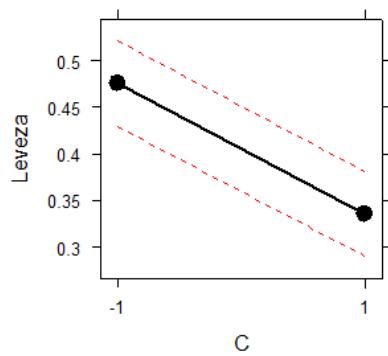


Gráfico de Efeitos

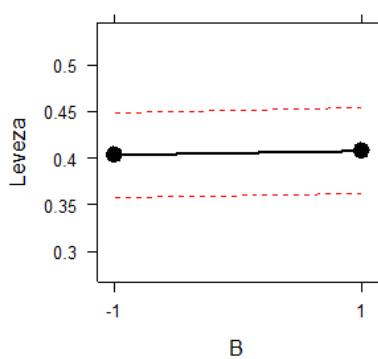


Gráfico de Interações

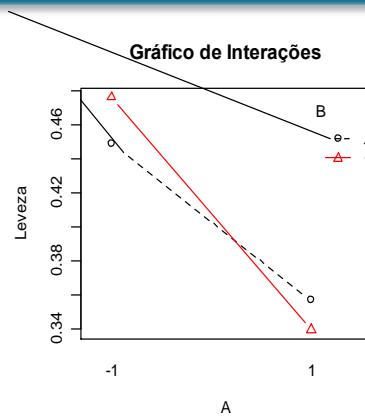


Gráfico de Interações

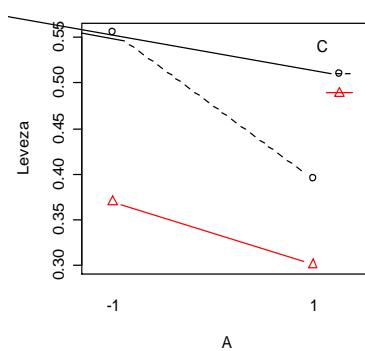
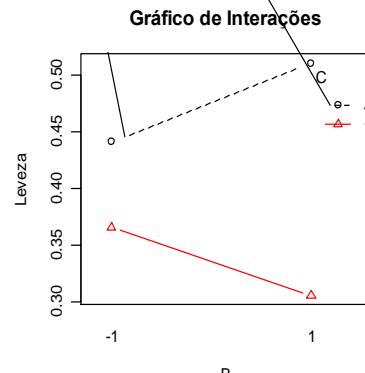


Gráfico de Interações



Exemplo

Exemplo 4.1.1.1: Uma empresa fabricante de bolos quer reduzir as reclamações relacionadas a dureza de sua massa. Para isso delineou um experimento 2^3 , com duas réplicas, com os fatores definidos da seguinte forma:

- A - Açúcar;
- B - Leite;
- C - Fermento.

Tabela da Anova

Fatores	G.L.	Soma de Quadrados	Quadrado Médio	Estat. F	P-valor
A	1	0,052326563	0,052326563	29,93403411	0,000593604
B	1	9,50625E-05	9,50625E-05	0,054381637	0,821463169
C	1	0,078540063	0,078540063	44,9297794	0,00015224
A:B	1	0,001958063	0,001958063	1,120133004	0,320797156
A:C	1	0,008235563	0,008235563	4,711251743	0,061766346
B:C	1	0,016835063	0,016835063	9,630698273	0,01459188
A:B:C	1	0,001785062	0,001785062	1,021166291	0,341826332
Residuals	8	0,0139845	0,001748063		

Os resultados mostrados na Tabela confirmam os indícios verificados na análise gráfica, que os fatores A e C e as interações AC e BC são significativas ao nível $\alpha = 0,05$, ou seja, eles influenciam na dureza da massa.

Coeficientes

Preditor	Efeitos	Estimativa	Desvio Padrão	Estat. T	P-valor
Intercepto		0,4055625	0,010452459	38,80067712	2,13903E-10
A	-0,114375	-0,0571875	0,010452459	-5,47120043	0,000593604
B	0,004875	0,0024375	0,010452459	0,233198707	0,821463169
C	-0,140125	-0,0700625	0,010452459	-6,70296795	0,00015224
A:B	-0,022125	-0,0110625	0,010452459	-1,05836336	0,320797156
A:C	0,045375	0,0226875	0,010452459	2,170541809	0,061766346
B:C	-0,064875	-0,0324375	0,010452459	-3,10333664	0,01459188
A:B:C	0,021125	0,0105625	0,010452459	1,010527729	0,341826332

Experimentos Fatoriais em Blocos

Dois Blocos

Além dos delineamentos já discutidos é possível construir um experimento factorial do tipo 2^k dividido em quatro blocos com $2^k/4=2^{k-2}$ observações em cada bloco.

Exemplo

- ❑ Considere um experimento factorial $2^5 = 32$ corridas;
- ❑ Para fazer 4 blocos, cada um terá 8 corridas;
- ❑ Escolhe-se dois efeitos para serem confundidos com os blocos, neste caso ADE e BCE.
- ❑ Se considerar apenas o ADE e BCE como colunas existirão exatamente quatro combinações de sinais diferentes para as 32 observações.

	Bloco 1		Bloco 2		Bloco 3		Bloco 4	
ADE	-		+		-		+	
BCE	-		-		+		+	
	0	abc	a	bc	b	abce	e	abcde
	ad	ace	d	abde	abd	ae	ade	bd
	bc	cde	abc	ce	c	bcde	bce	ac
	abcd	bde	bcd	acde	acd	de	ab	cd

O efeito geral de ADE e BCE é definido como o produto entre eles

$$(ADE)(BCE) = ABCDE^2 = ABCD$$

(já que E² = 1 sempre) e então ABCD também é confundido com os blocos.

Experimentos Fatoriais em Blocos

Quatro ou mais Blocos

Apresentamos alguns exemplos de efeitos que podem ser confundidos.

$$(ABEF)(ABCD) = A^2B^2CDEF = CDEF$$

$$(ABEF)(ACE) = A^2BCE^2F = BCF$$

$$(ABCD)(ACE) = A^2BC^2ED = BDE$$

$$(ABEF)(ABCD)(ACE) = A^3B^2C^2DE^2F = ADF$$

A Tabela mostra a estrutura de blocos para Experimentos Fatoriais 2^k

Número de fatores, k	Número de blocos, 2^p	Tamanho do bloco, 2^{k-p}	Efeito escolhido	Interações confundidas
3	2	4	ABC	ABC
	4	2	AB,AC	AB,AC,BC
4	2	8	ABCD	ABCD
	4	4	ABC,ACD	ABC,ACD,BD
	8	2	AB,BC,CD	AB,BC,CD,AC,BD,AD,ABCD
5	2	16	ABCDE	ABCDE
	4	8	ABC,CDE	ABC,CDE,ABDE
	8	4	ABE,BCE,CDE	ABE,BCE,CDE,AC,ABCD,BD,ADE
	16	2	AB,AC,CD,DE	Todos os fatores de 2º grau, 4º grau (15 interações)
6	2	32	ABCDEF	ABCDEF
	4	16	ABC,F,CDEF	ABC,F,CDEF,ABDE
	8	8	ABEF,ABC,D,ACE	ABEF,ABC,D,ACE,BC,F,BD,E,CDEF,ADF
	16	4	AB,F,AC,F,BD,F,DEF	ABE,ACE,BDF,DEF,BC,ABCD,ABDE,AD,ACDE CE,BDF,BCDEF,ABCE,F,AE,F,BE
7	32	2	AB,BC,CD,DE,EF	Todos os fatores de 2º grau, 4º grau, 6º grau (31 interações)
	2	64	ABCDEFG	ABCDEFG
	4	32	ABC,F,G,CDEF,G	ABC,F,G,CDEF,G,ABDE
	8	16	ABC,DEF,AFG	ABC,DEF,AFG,ABCDEF,BCFG,ADEG,BCDEG
	16	8	ABCD,EFG,CDE,ADG	ABCD,EFG,CDE,ADG,ABCDEF,G,ABE,BCG CDFG,ADEF,ACEG,ABFG,BCEF,BDEG,AC,F,BDF
	32	4	ABG,BCG,CDG,DEG,EFG	ABG,BCG,CDG,DEG,EFG,AC,BD,CE,DF,AE,BE,ABCD ABDE,ABEF,BCDE,BCEF,CDEF,ABDEFG,ADG,ACDEG,ACEFG ABDFG,ABCEG,BEG,BDEFG,CFG,ADEF,ACDF,ABC,F,AFG
64	2	AB,BC,CD,DE,EF,FG	Todos os fatores de 2º grau, 4º grau, 6º grau (63 interações)	

Experimentos Fatoriais Fracionados

Experimentos Fatoriais Fracionados

Experimentos Fatoriais Fracionados



Os experimentos fatoriais fracionados estão entre as ferramentas mais usadas nas indústrias no mundo.

Estes experimentos fatoriais estão fundamentalmente baseados em três ideias:

Efeitos principais



Quando temos muitas variáveis, o sistema ou processo será dirigido por alguns efeitos principais e interações de baixa ordem.

Propriedades projetivas



Fatoriais fracionados podem ser projetados para que os fatores mais significantes tenham os maiores efeitos.

Experimentos sequenciais



É possível combinar as corridas de dois (ou mais) fatoriais fracionados para montar sequencialidade em delineamentos grandes para estimar os fatores e interações de interesse.

Experimentos Fatoriais Fracionados

Meia fração de um experimento fatorial fracionado

Temos um experimento com três fatores com dois níveis cada;

Serão executados metade dos 8 tratamentos;

Selecionemos quatro tratamentos a,b,c e abc; com sinal positivo na coluna ABC, como nossa meia fração. ABC é chamado de gerador desta fração.

A Tabela representa um delineamento do tipo 2^3 . Suponha que neste caso selecionemos quatro tratamentos a, b, c e abc como nossa meia fração. Note que neste caso temos ainda as interações de segunda ordem, mas escolhemos definir como a meia fração os termos que tem sinal positivo na coluna ABC.

Combinação do Tratamento	Efeito Fatorial							
	I	A	B	C	AB	AC	BC	ABC
a	+	+	-	-	-	-	+	+
b	+	-	+	-	-	+	-	+
c	+	-	-	+	+	-	-	+
abc	+	+	+	+	+	+	+	+
ab	+	+	+	-	+	-	-	-
ac	+	+	-	+	-	+	-	-
bc	+	-	+	+	-	-	+	-
(0)	+	-	-	-	+	+	+	-

Experimentos Fatoriais Fracionados

Meia fração de um experimento fatorial fracionado

As combinações do tratamento no delineamento fatorial 2^{3-1} rendem três graus de liberdade para estimar os efeitos principais. Que podem ser definidos como:

$$A = \frac{1}{2}(a - b - c + abc)$$

$$B = \frac{1}{2}(-a + b - c + abc)$$

$$C = \frac{1}{2}(-a - b + c + abc)$$

Para saber se o fator A está confundido com algum outro fator, podemos fazer o cálculo:

$$\begin{aligned} A &= A * I = A * ABC = A^2 * BC \quad (A^2 \text{ é sempre } 1) \\ A &= BC \end{aligned}$$

Temos que os efeitos das interações de 2^a ordem são dados por:

$$BC = \frac{1}{2}(a - b - c + abc)$$

$$AC = \frac{1}{2}(-a + b - c + abc)$$

$$AB = \frac{1}{2}(-a - b + c + abc)$$

Podemos verificar, $A = BC$, $B = AC$, $C = AB$. Assim é impossível diferenciar as relações entre A e BC, B e AC, e C e AB.

Exemplo

Se considerarmos $A = A + BC$ e $A = A - BC$

Esta operação implicaria em:

$$\frac{1}{2}(A + A) = \frac{1}{2}(A + BC + A - BC) = A$$

e

$$\frac{1}{2}(A - A) = \frac{1}{2}(A + BC - A + BC) = BC$$

Portanto, todas as três partes das combinações lineares são demonstradas da seguinte maneira:

	$\frac{1}{2}(A + A)$	$\frac{1}{2}(A - A)$
A	A	BC
B	B	AC
C	C	AB

Experimentos Fatoriais Fracionados

Resolução do tipo III



Estas são as resoluções nos quais nenhum efeito principal está relacionado com qualquer outro efeito principal, mas podem estar relacionado com interações de segunda ordem que por sua vez podem estar relacionadas também com interações de segunda ordem ou de ordem superior.

Resolução do tipo IV



Esta resolução determina que nenhum efeito principal está relacionado com outro efeito principal ou então interação de segunda ordem, mas as interações de segunda ordem estão relacionadas entre si. Um planejamento 2^{4-1} com $I = ABCD$ segue resolução do tipo IV (2^{5-1}_V).

Resolução do tipo V



Estes são os experimentos em que nem os efeitos principais ou as interações de segundo grau estão relacionados com os efeitos principais ou as interações de segundo grau, mas as interações de segundo grau podem estar relacionadas com interações de terceiro ou maior grau. Um planejamento com segue resolução do tipo V .

Podemos entender este procedimento da seguinte forma:

	Fatorial Completo		$I = ABC$			$I = -ABC$		
Corrida	A	B	A	B	C=AB	A	B	C=-AB
1	-	-	-	-	+	-	-	-
2	+	-	+	-	-	+	-	+
3	-	+	-	+	-	-	+	+
4	+	+	+	+	+	+	+	-

Experimentos Fatoriais Fracionados

Fatoriais Fracionados Sequenciais

Usando planejamentos fatoriais fracionados podemos ter uma grande economia e eficiência no experimento, particularmente se o experimento for conduzido sequencialmente.

Por exemplo, suponha que tenhamos que investigar um fatorial $k = 4$ fatores ($2^4 = 16$). Porém é preferível termos um experimento pela metade, ou seja, 2_{IV}^{4-1} fatorial com 8 corridas, analisar os resultados e então decidir qual o melhor conjunto para correr em seguida.

Se for necessário desenvolver ambiguidades é preferível que o experimento seja conduzido de maneira alternada até completar o fatorial. Quando este método for escolhido, ambas as frações do fatorial representam blocos do planejamento completo, com a maior ordem das interações confundida com os blocos (ABCD).

Podemos afirmar que os experimentos sequenciais podem perder informação, mas apenas nas interações mais altas.

Podemos também aprender o suficiente com a primeira fração e poderemos fazer novas análises dos dados, excluir fatores, interações e até aumentar o intervalo de varredura de certas variáveis.

Exemplo

Exemplo 5.2.1: Considere a taxa de filtragem definida na tabela a seguir. A taxa de filtragem depende dos fatores envolvidos no processo:

- Temperatura (A);
- Pressão (B);
- Concentração de gás (utilizado para esterilização) (C);
- Taxa de filtragem (D).

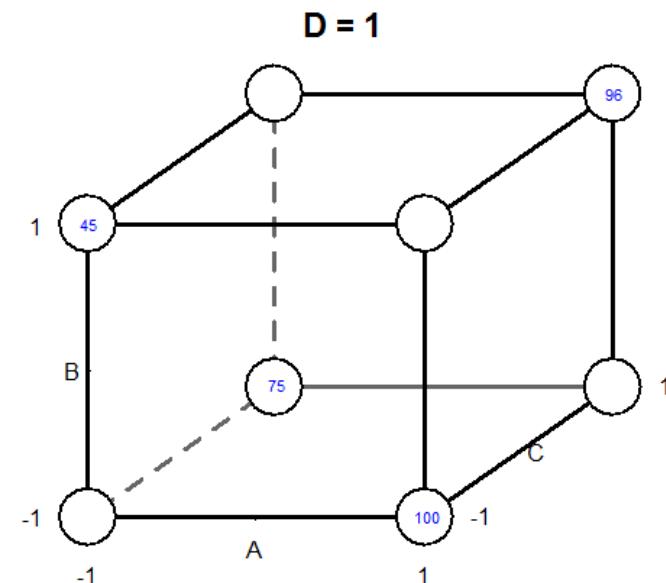
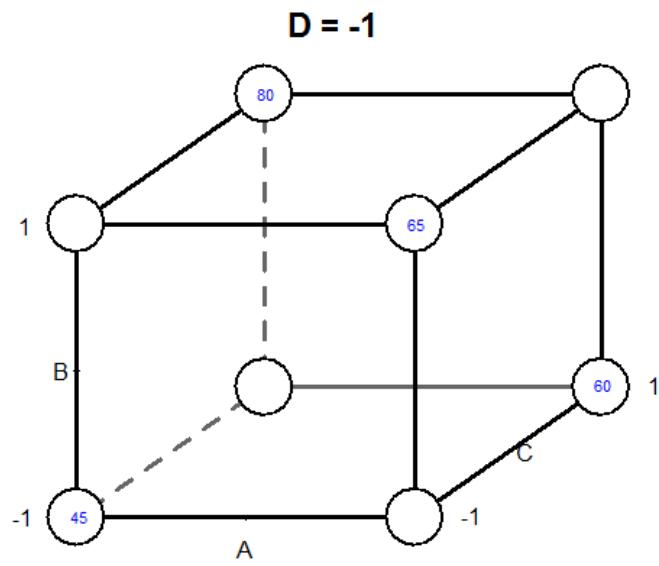
Corridas	Planejamento básico			D=ABC	Combinação do tratamento	Taxa de filtragem
	A	B	C			
1	-	-	-	-	(0)	45
2	+	-	-	+	ad	100
3	-	+	-	+	bd	45
4	+	+	-	-	ab	65
5	-	-	+	+	cd	75
6	+	-	+	-	ac	60
7	-	+	+	-	bc	80
8	+	+	+	+	abcd	96

Exemplo

Exemplo 5.2.1: Considere a taxa de filtragem definida na tabela a seguir. A taxa de filtragem depende dos fatores envolvidos no processo:

- Temperatura (A);
- Pressão (B);
- Concentração de gás (utilizado para esterilização) (C);
- Taxa de filtragem (D).

A	B	C	D	Filtragem
-1	-1	-1	-1	45
1	-1	-1	-1	
-1	-1	1	-1	
1	-1	1	-1	60
-1	1	-1	-1	
1	1	-1	-1	65
-1	1	1	-1	80
1	1	1	-1	
-1	-1	-1	1	
1	-1	-1	1	100
-1	-1	1	1	75
1	-1	1	1	
-1	1	-1	1	45
1	1	-1	1	
-1	1	1	1	96
1	1	1	1	



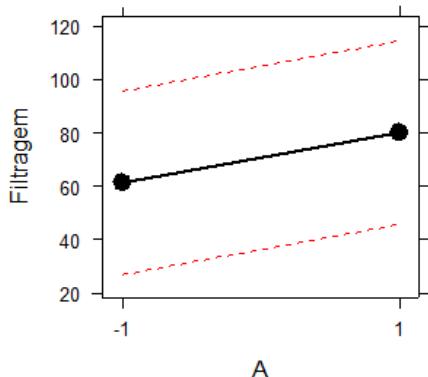
Número de fatores (k)	Fração	Número de Corridas	Geradores do Planejamento
3	2_{III}^{3-1}	4	$C = \pm AB$
4	2_{IV}^{4-1}	8	$D = \pm ABC$
5	2_V^{5-1}	16	$E = \pm ABCD$
	2_{III}^{5-2}	8	$D = \pm AB$
			$E = \pm AC$
6	2_{IV}^{6-1}	32	$F = \pm ABCDE$
	2_{IV}^{6-2}	16	$E = \pm ABC$
			$F = \pm BCD$
	2_{III}^{6-3}	8	$D = \pm AB$
			$E = \pm AC$
			$F = \pm BC$
7	2_{VII}^{7-1}	64	$G = \pm ABCDEF$
	2_{IV}^{7-2}	32	$E = \pm ABC$
			$G = \pm ABDE$
	2_{IV}^{7-3}	16	$E = \pm ABC$
			$F = \pm BCD$
			$G = \pm ACD$
	2_{III}^{7-4}	8	$D = \pm AB$
			$E = \pm AC$
			$F = \pm BC$
			$G = \pm ACD$
8	2_V^{8-2}	64	$G = \pm ABCD$
			$H = \pm ABEF$
	2_{IV}^{8-3}	32	$F = \pm ABC$
			$G = \pm ABD$
			$H = \pm BCDE$
	2_{IV}^{8-4}	16	$E = \pm BCD$
			$F = \pm ACD$
			$G = \pm ABC$
			$H = \pm ABD$

Exemplo

Exemplo 5.2.1: Considere a taxa de filtragem definida na tabela a seguir. A taxa de filtragem depende dos fatores envolvidos no processo:

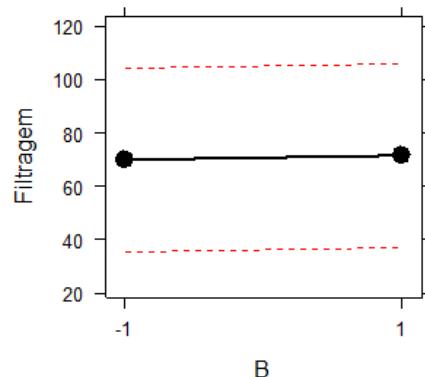
- Temperatura (A);
- Pressão (B);
- Concentração de gás (utilizado para esterilização) (C);
- Taxa de filtragem (D).

Gráfico de Efeitos



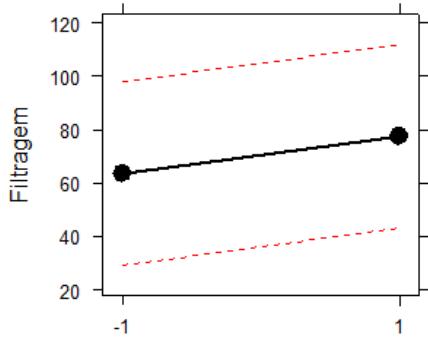
A

Gráfico de Efeitos



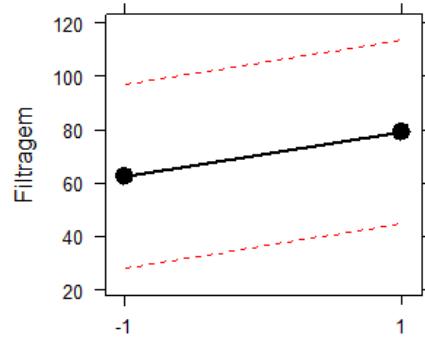
B

Gráfico de Efeitos



C

Gráfico de Efeitos



D

		Efeito	Límite Superior
A	<i>Límite Inferior</i>	61,25	95,72854161
-1	26,77145839		
1	45,77145839	80,25	114,7285416
B	<i>Límite Inferior</i>	70	104,4785416
-1	35,52145839		
1	37,02145839	71,5	105,9785416
C	<i>Límite Inferior</i>	63,75	98,22854161
-1	29,27145839		
1	43,27145839	77,75	112,2285416
D	<i>Límite Inferior</i>	62,5	96,97854161
-1	28,02145839		
1	44,52145839	79	113,4785416

Experimentos Fatoriais Fracionados

Um quarto de fração de um planejamento 2^k

Considere um planejamento 2^{6-2} , suponha que escolhamos

$$I = ABCE \text{ e } I = BCDF$$

como os geradores.

Matricialmente podemos descrever o experimento como:

Corridas	A	B	C	D	$E = ABC$	$F = BCD$
1	-	-	-	-	-	-
2	+	-	-	-	+	-
3	-	+	-	-	+	+
4	+	+	-	-	-	+
5	-	-	+	-	+	+
6	+	-	+	-	-	+
7	-	+	+	-	-	-
8	+	+	+	-	+	-
9	-	-	-	+	-	+
10	+	-	-	+	+	+
11	-	+	-	+	+	-
12	+	+	-	+	-	-
13	-	-	+	+	+	-
14	+	-	+	+	-	-
15	-	+	+	+	-	+
16	+	+	+	+	+	+

Experimentos Fatoriais Fracionados

Um quarto de fração de um planejamento 2^k

A interação entre os geradores ABCE e BCDF é ADEF, portanto a relação completa para o experimento é:

$$I = ABCE = BCDF = ADEF$$

Por exemplo, para encontrar os aliases de A podemos fazer:

- $AABCE = AABCE = BCE$, pois $AA = 1$ sempre (A só assume os valores +1 ou -1),
- $ABCDF = ABCDF$;
- $AADEF = AADEF = DEF$, novamente pois $AA = 1$.

Assim, vemos que quando estimarmos A, estaremos estimando:

$$A + BCE + DEF + ABCDF$$

A estrutura completa de aliases para este caso é mostrada na seguinte tabela:

$A = BCE = DEF = ABCDF$	$AD = EF = ABCF = BCDE$
$B = ACE = CDF = ABDEF$	$AE = BC = DF = ABCDEF$
$C = ABE = BDF = ACDEF$	$AF = DE = ABCD = BCEF$
$D = AEF = BCF = ABCDE$	$BD = CF = ABEF = ACDE$
$E = ABC = ADF = BCDEF$	$BF = CD = ABDE = ACEF$
$F = ADE = BCD = ABCEF$	$ABD = ACF = BEF = CDE$
$AB = CE = ACDF = BDEF$	$ABF = ACD = BDE = CEF$
$AC = BE = ABDF = CDEF$	

Exemplo

Exemplo 5.1.1.1: Um experimentador quer avaliar o comportamento do desgaste do pinhão do motor de uma máquina colheitadeira de cana, em relação a quatro fatores: abertura da lateral do chassi, janela da caneca, formato do disco e formato do rolo. O objetivo é determinar qual configuração estabelece um menor desgaste do pinhão. Utilizando 8 máquinas do mesmo modelo, mesmo ano de fabricação e utilizando a mesma frente de trabalho, realizamos um experimento factorial fracionado 2^{4-1} com três réplicas. As unidades experimentais foram de 250 horas de trabalho por réplica. Os dados são dados na tabela a seguir.

Lateral chassi (LC)	Janela canela (JC)	Formato disco (FD)	Rolo (R)	Desgaste
menor	com janela	sem recorte	com furo	0,2233
menor	sem janela	com recorte	com furo	0,1233
menor	sem janela	com recorte	com furo	0,1300
maior	sem janela	com recorte	sem furo	0,1867
menor	com janela	sem recorte	com furo	0,2533
menor	sem janela	com recorte	com furo	0,1700
menor	com janela	com recorte	sem furo	0,2400
menor	sem janela	sem recorte	sem furo	0,3033
maior	com janela	com recorte	com furo	0,1933
maior	com janela	sem recorte	sem furo	0,2500
maior	com janela	com recorte	com furo	0,1900
maior	com janela	sem recorte	sem furo	0,2533
maior	sem janela	sem recorte	com furo	0,2067
menor	com janela	com recorte	sem furo	0,2033
maior	com janela	com recorte	com furo	0,1767
menor	sem janela	sem recorte	sem furo	0,2200
menor	sem janela	sem recorte	sem furo	0,2133
maior	sem janela	com recorte	sem furo	0,1800
maior	com janela	sem recorte	sem furo	0,2333
maior	sem janela	com recorte	sem furo	0,1967
maior	sem janela	sem recorte	com furo	0,2100
maior	sem janela	sem recorte	com furo	0,2033
menor	com janela	sem recorte	com furo	0,2037
menor	com janela	com recorte	sem furo	0,1967

Aplicações

Aplicação 1: Uma empresa está interessada em minimizar a granulometria de material residual indesejado em certo produto na produção do mesmo. Para isto, foi feito um planejamento de experimento no qual foram selecionados as variáveis:

Damper – mede o controle do fluxo de material;

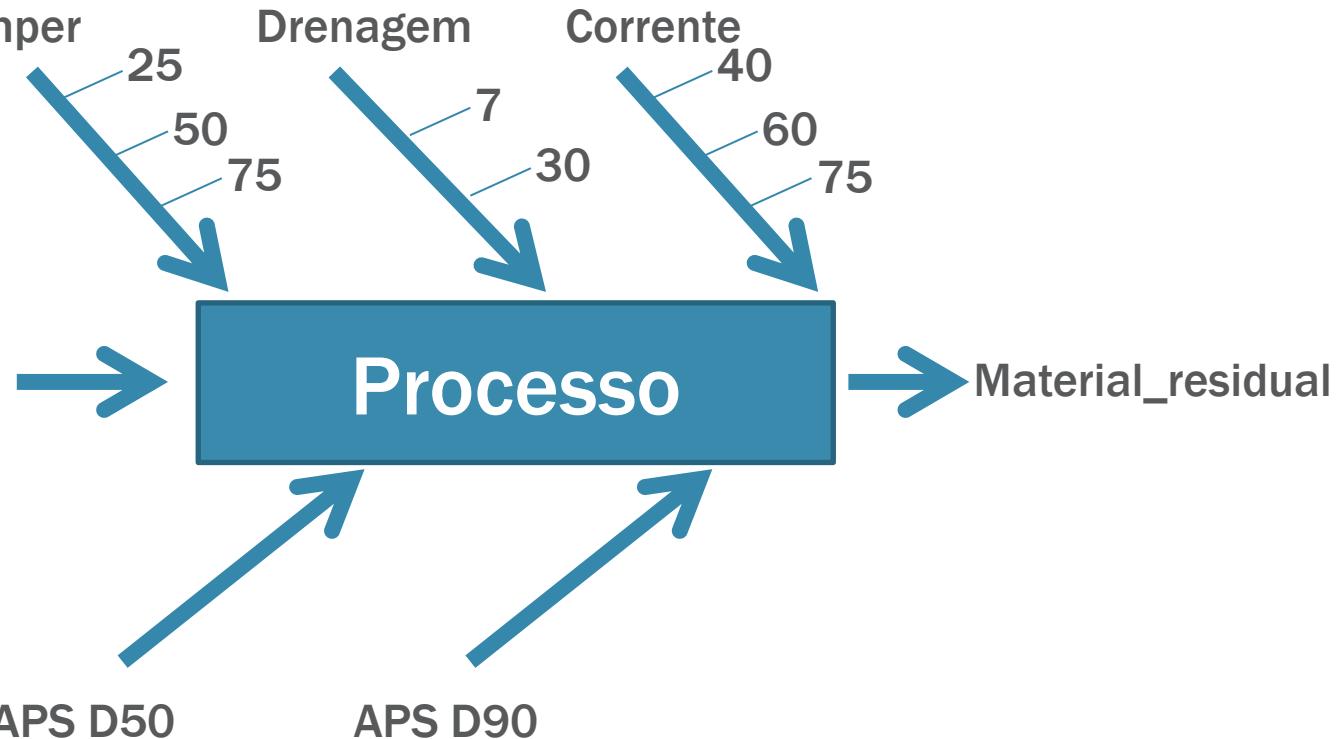
Drenagem – representa a quantidade de dias para a secagem do material;

Corrente – é o fluxo de ar do duto;

APS D90 – representa a distribuição do tamanho de partículas do resíduo dadas em (%);

APS D50 – representa a distribuição do tamanho de partículas do resíduo dadas em (%);

Material_residual – é a quantidade (gramas) de resíduos de materiais indesejados que ficaram retidos na peneira.



Objetivo

Otimizar a quantidade mínima de flocos de resíduos indesejados na peneira (próximo de zero).

As variáveis explicativas são Damper, Drenagem, Corrente, APS D90 e APS D50.

A variável resposta é a **material_residual**.

Aplicação 2: Uma empresa deseja testar o efeito de duas diferentes dietas no desempenho de tilápias na engorda. Os peixes estão divididos aleatoriamente em 40 gaiolas e foram submetidos a dieta durante 90 dias.

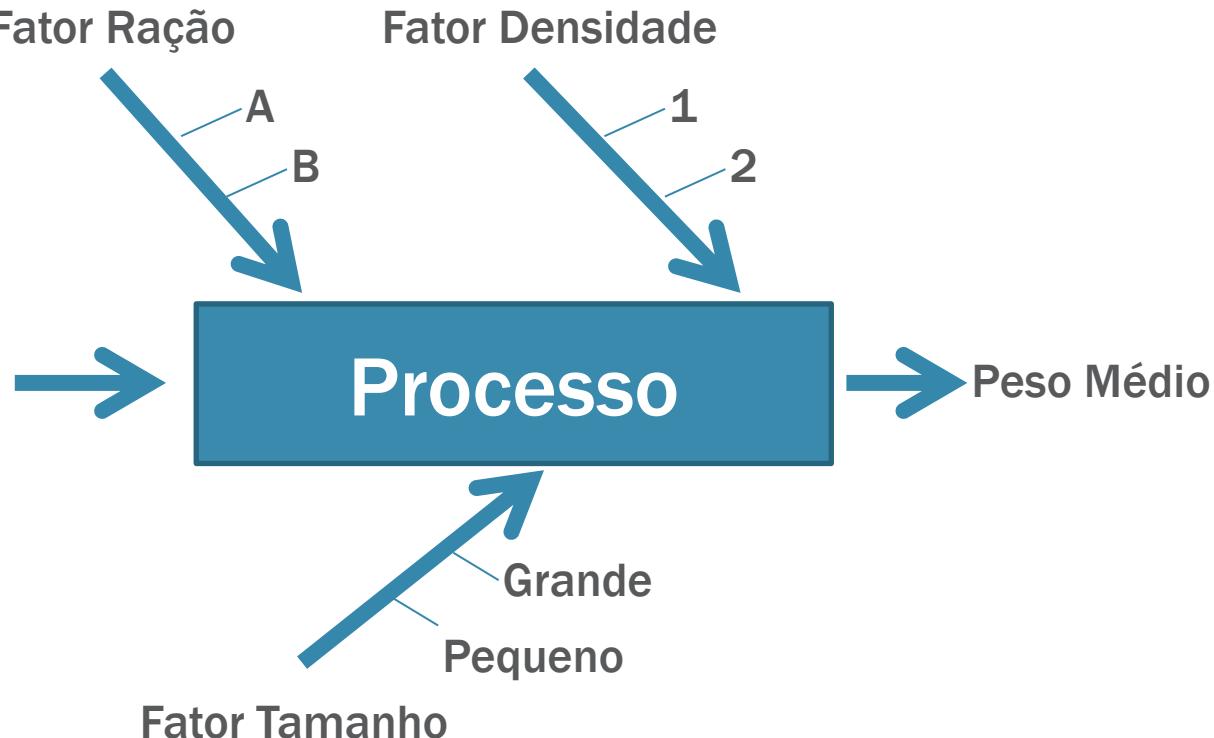
Para este problema foram considerados 3 fatores:

Ração: Ração A e ração B

Tamanho : Peixes pequenos e peixes grandes

Densidade: Gaiolas com 102 (1) e 150 (2) peixes

As gaiolas foram montadas com duas quantidades de peixes, **densidade 1** - 102 peixes; **densidade 2** - 150 peixes. Os peixes foram colocados nas gaiolas aleatoriamente.



Objetivo

Testar o efeito de duas diferentes dietas no desempenho de tilápias na engorda.

As variáveis explicativas são Ração, Tamanho e Densidade.

A variável resposta é o **peso médio** por peixe ganho durante os 90 dias em que foram submetidos à dieta.

Regressão Logística

Modelo Logístico

Suponha uma amostra de n observações independentes da terna (x_i, m_i, y_i) , $i = 1, 2, \dots, n$, no qual:



- x_i é o valor da variável explicativa;
- m_i é a quantidade de replicatas (número de ensaios);
- y_i número de replicatas detectada com microrganismos em m_i replicatas;
- n é o total de combinações.

Com isso, assumimos que a variável resposta tem distribuição de probabilidade binomial tal que:

$$P[Y_i = y_i] = \binom{m_i}{y_i} \pi_i^{y_i} (1 - \pi_i)^{m_i - y_i}$$

Modelo Logístico

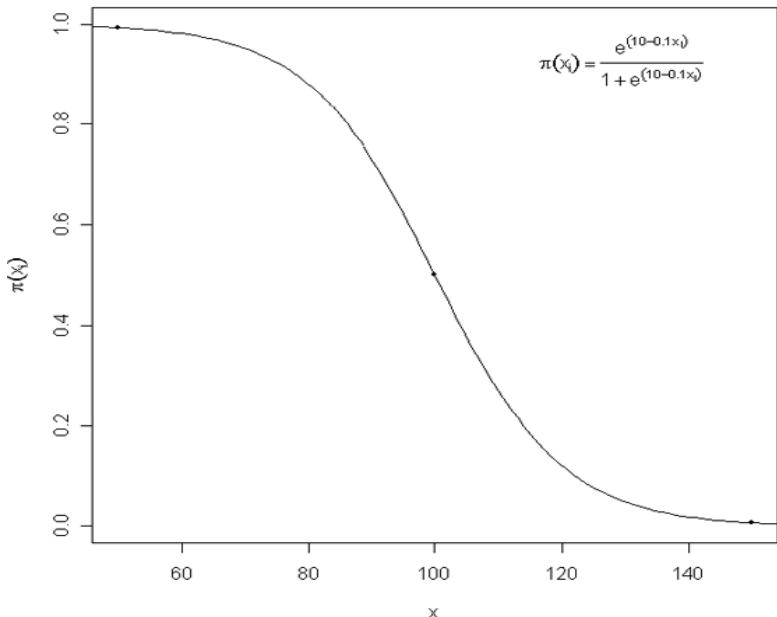
Para adequarmos a resposta média ao modelo linear usamos a função de ligação, que é dada por:

$$\pi_i = \pi(x_i) = \frac{e^{\beta_0 + \beta_1 x_i}}{1 + e^{\beta_0 + \beta_1 x_i}}, \quad "i = 1, \dots n"$$

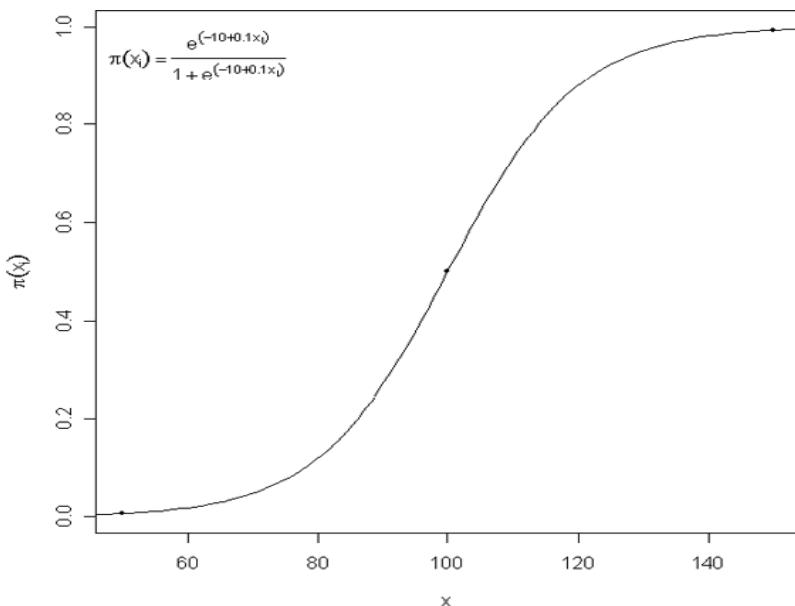
A função de ligação também pode ser escrita como:

$$g(X) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 x_i$$

Modelo Logístico

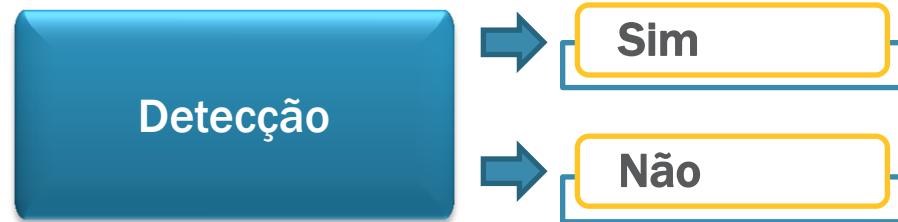


Modelo Logístico com β negativo

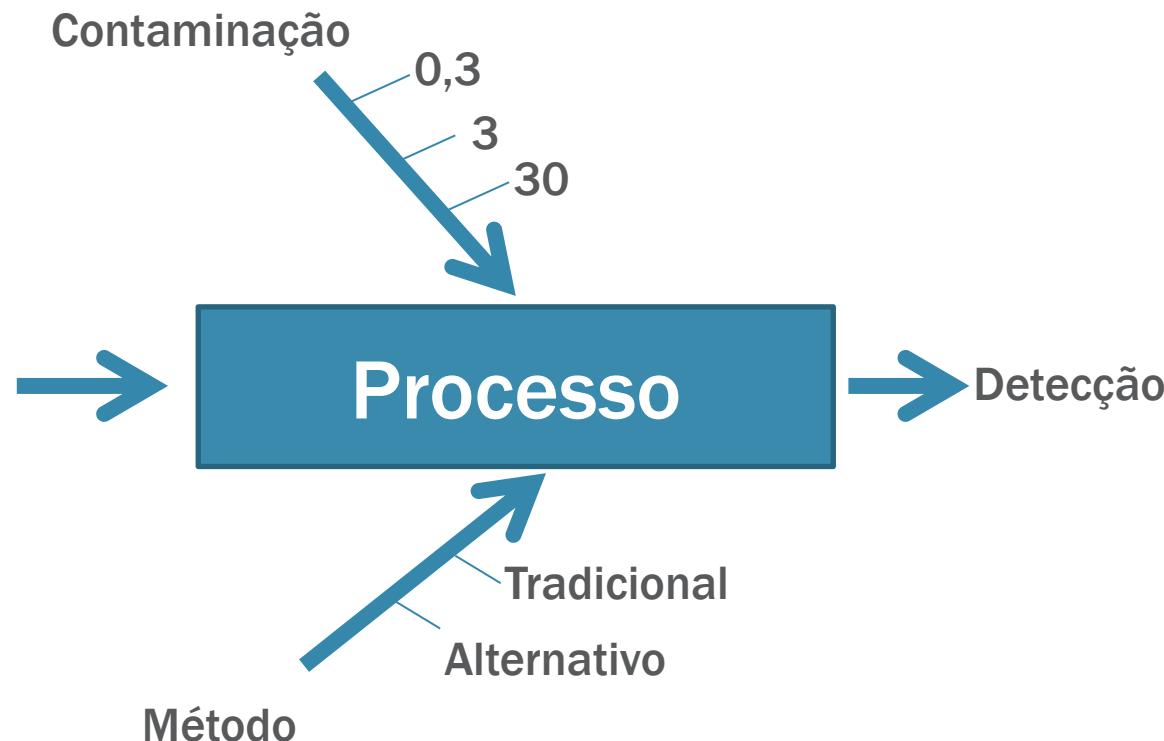


Modelo Logístico com β positivo

Aplicação 3: Uma empresa farmacêutica tem interesse em avaliar o limite de detecção de microrganismos, para isto realizou-se um experimento no qual foram considerados:



- Nove microrganismos;
- Três lotes de produto;
- Sete repetições em cada combinação.



Objetivo

Avaliar o limite de detecção.

As variáveis explicativas são Contaminação e Método.

A variável resposta é a **Detecção**, uma variável binária que diz se foi detectado ou não.

Aplicação 4: Em uma pesquisa de satisfação, 50 clientes opinaram sobre 4 programas de recompensas de cartão de crédito emitindo uma nota para cada programa. Queremos saber quais programas têm maior aceitação dos clientes.

Os programas avaliados foram:

1. Programa de incentivo a viagens: compras viram pontos e dão direitos a passagens aéreas;
2. Programa de incentivo a compras: pontos acumulados que se transformam em compras grátis em grandes redes comerciais;
3. Programa de prêmios: clientes concorrem a prêmios semanais e mensais em dinheiro;
4. Programa de incentivo ao carro 0km: as compras transformam-se em descontos na compra de um carro 0km.

Para este problema consideramos duas variáveis:

Programa: 1, 2, 3 e 4.

Nota: de 5 a 10.

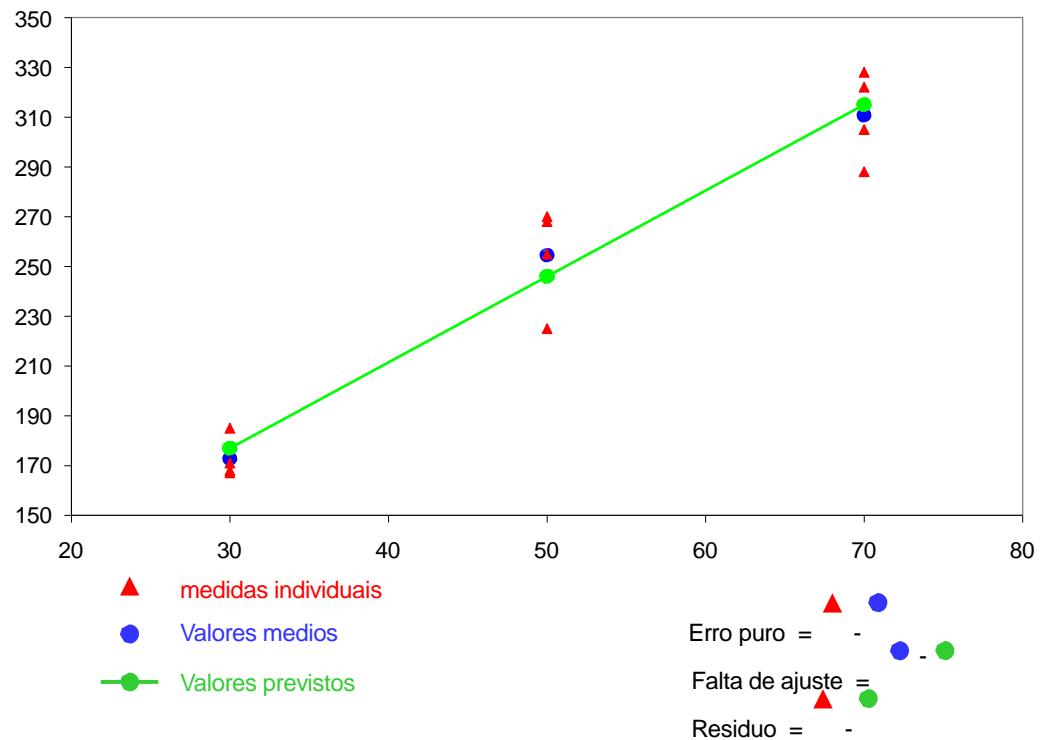
Objetivo

Identificar qual programa tem maior aceitação do cliente.

A variável explicativa é Programa.

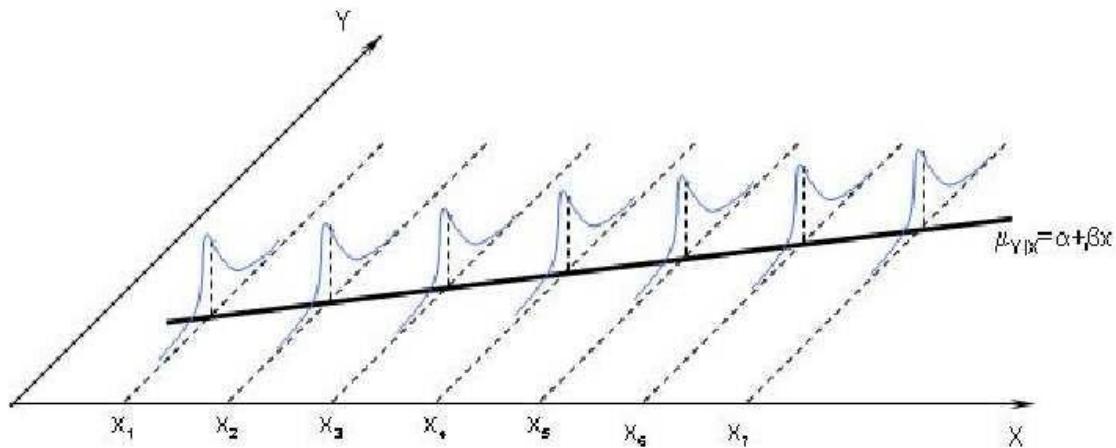
A variável resposta é a **Nota**.

Erro puro

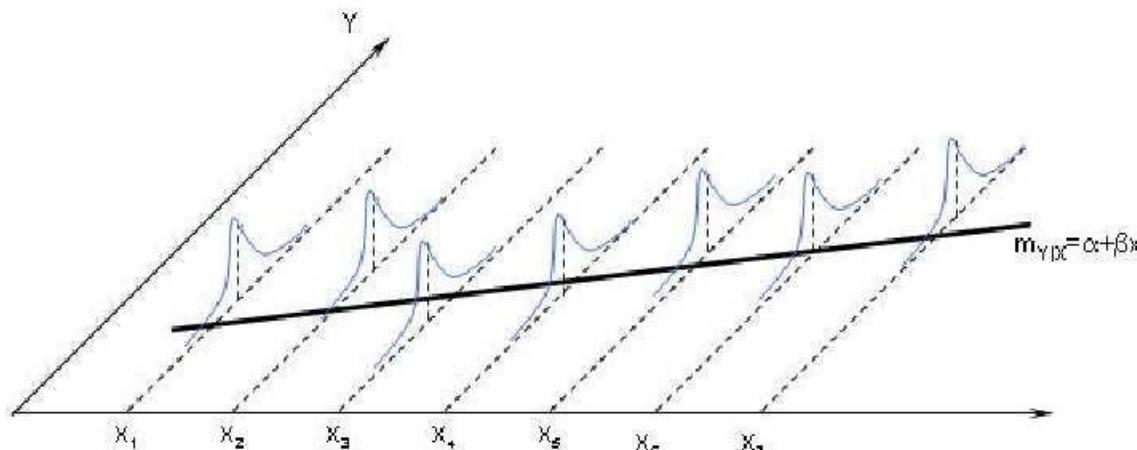


Teste da Falta de Ajuste (Lack of Fit)

Após o ajuste, é importante verificar se o modelo linear é adequado.



Reta de regressão
perfeitamente ajustada
sem Falta de Ajuste



Reta de regressão com
Falta de Ajuste

Teste da falta de ajuste

O teste consiste em comparar através de uma ANOVA a variância da falta de ajuste com a variância do erro puro conforme abaixo:

A analise é feita conforme segue:

Se valor $p \geq 0,05$ não há falta de ajuste – a curva é adequada para o modelo estudado

Se valor $p < 0,05$ há falta de ajuste – a curva não é adequada para o modelo estudado