

Estatística inferencial no *software* R

Por meio do pacote Rcmdr

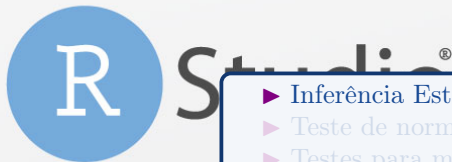
Diogo Macedo Mendes

Keyla Megumi Sano de Oliveira

Profa. Dra. Giovana Fumes Ghantous

December 4, 2023





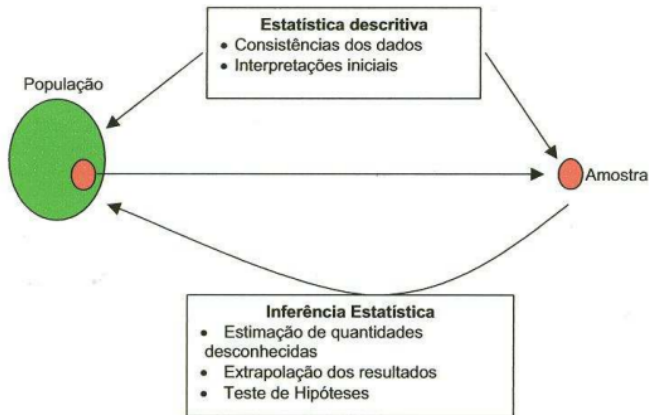
- ▶ Inferência Estatística
- ▶ Teste de normalidade
- ▶ Testes para médias
- ▶ Teste Qui-quadrado
- ▶ Testes de Correlação
- ▶ Regressão Linear Simples



É o processo de **tirar conclusões** ou **fazer previsões** sobre uma população com base em informações obtidas de uma amostra desta população.

Inferência - Conceitos Básicos

Etapas da análise estatística.



Inferência - Intervalo de confiança $\gamma = (1 - \alpha)$



É uma faixa de valores que é usada para **estimar um parâmetro desconhecido de uma população com um certo grau de confiança.**

Em termos simples, um intervalo de confiança fornece um intervalo de valores dentro do qual é provável que o valor real do parâmetro esteja, com base nos dados da amostra.

Por exemplo, se você calcular um intervalo de confiança de 95% para a média de uma população, isso significa que há 95% de confiança de que o intervalo contenha a verdadeira média da população. Quanto maior o nível de confiança (por exemplo, 95% em vez de 90%), o intervalo de confiança será mais amplo, refletindo uma maior incerteza.

É um método estatístico usado para avaliar a validade de uma afirmação (hipótese) sobre uma população com base em uma amostra. A hipótese de interesse é chamada de **hipótese nula** (H_0), e a outra de **hipótese alternativa** (H_a).

- Por exemplo: Deseja-se saber se a média da distribuição de notas de uma turma é **igual** a 8 ou não, as hipóteses são dadas por:

$$H_0 : \mu = 8$$

$$H_a : \mu \neq 8.$$

Inferência - Teste de Hipóteses

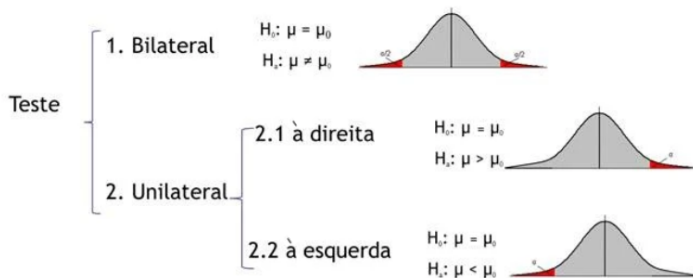


Table: Relação entre os erros tipo I e tipo II no processo de decisão.

	Não rejeitar H_0	Rejeitar H_0
H_0 verdadeira	Decisão correta	Erro do tipo I (α)
H_0 falsa (ou H_a)	Erro do tipo II (β)	Decisão correta

Inferência - Teste de Hipóteses

- Os testes podem ser bilaterais (bicaudais) ou unilaterais (monocaudais).
- Por exemplo, as hipóteses de testes para médias populacionais podem ser da forma:





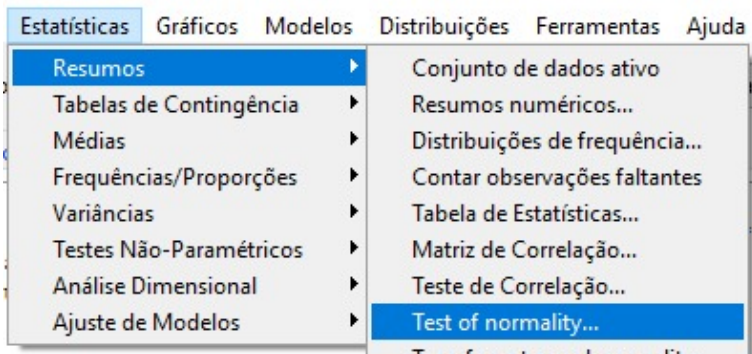
Studio®

- ▶ Inferência Estatística
- ▶ Teste de normalidade
- ▶ Testes para médias
- ▶ Teste Qui-quadrado
- ▶ Testes de Correlação
- ▶ Regressão Linear Simples



Inferência - Teste de normalidade

Para avaliar se a variável altura do banco de dados "AULA2" segue uma distribuição normal, os seguintes passos podem ser utilizados: **Estatísticas > Resumos > Test of normality**

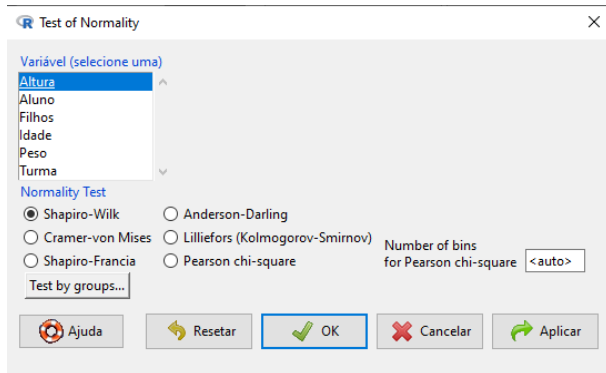


Inferência - Teste de normalidade

Neste caso, as hipóteses do teste são:

H_0 : Os dados seguem distribuição normal

H_a : Os dados não seguem distribuição normal



The image shows the 'Test of Normality' dialog box in R. The window title is 'R Test of Normality'. It contains a list of variables under the heading 'Variável (selecione uma)'. The variable 'Altura' is selected. Below this, under the heading 'Normality Test', there are several radio button options: 'Shapiro-Wilk' (selected), 'Anderson-Darling', 'Cramer-von Mises', 'Lilliefors (Kolmogorov-Smirnov)', 'Shapiro-Francia', and 'Pearson chi-square'. To the right of these options is a text field labeled 'Number of bins for Pearson chi-square' with the value '<auto>'. At the bottom left, there is a text field labeled 'Test by groups...'. At the bottom of the dialog are five buttons: 'Ajuda' (Help), 'Resetar' (Reset), 'OK' (highlighted with a blue border), 'Cancelar' (Cancel), and 'Aplicar' (Apply).

Inferência - Teste de normalidade

```
Output

> normalityTest(~Altura, test="shapiro.test", data=Dataset)

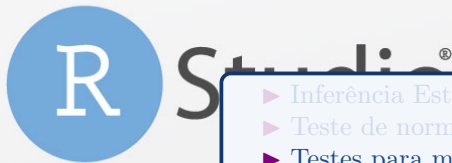
Shapiro-Wilk normality test

data:  Altura
W = 0.9279, p-value = 0.0875
```

H_0 = A Distribuição dos dados é normal

P-valor > 0,05, não rejeita-se H_0

Como p-valor > 0,05, não rejeita-se a hipótese nula. Portanto, conclui-se que a distribuição dos dados da variável altura seguem uma distribuição normal.



- ▶ Inferência Estatística
- ▶ Teste de normalidade
- ▶ Testes para médias
- ▶ Teste Qui-quadrado
- ▶ Testes de Correlação
- ▶ Regressão Linear Simples



Teste t para uma média

- O teste t para média de uma população deve ser usado quando deseja-se testar se a média é igual/menor ou igual/menor ou igual/menor ou igual/diferente a um valor especificado.
- Para realizar esse teste, é preciso partir do pressuposto que a distribuição dos dados é normal, e a variância populacional é desconhecida.

Teste t para uma média

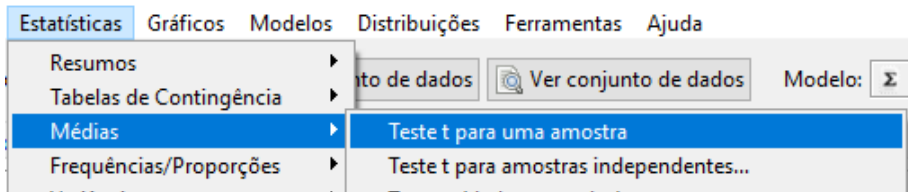
Exemplo. Com base no conjunto de dados do arquivo "AULA2" deseja-se testar se, ao nível de 5% de significância, a média das alturas dos alunos é igual à 1,62 metros. Suponha para realizar o teste que a amostra seja proveniente de uma distribuição normal.

$$H_0 : \mu = 1,62$$

$$H_a : \mu \neq 1,62$$

Teste t para uma média

Estatísticas > Médias > Teste t para uma amostra



Teste t para uma média

Teste-t para amostra simples

Variável (selecione uma)

Altura

Aluno

Filhos

Idade

Peso

Turma

Hipótese alternativa

☒ Média da População $\neq \mu_0$

Hipótese nula: $\mu =$

1.62

☐ Média da população $< \mu_0$

Nível de Confiança:

.95

☐ Média da população $> \mu_0$



Ajuda



Resetar



OK



Cancelar



Aplicar

Teste t para uma média

Output

```
> with(aula2, (t.test(Altura, alternative = "two.sided", mu = 1.62, conf.level = .95)))
```

One Sample t-test

```
data: Altura
t = 7.3179, df = 23, p-value = 0.000000191
alternative hypothesis: true mean is not equal to 1.62
95 percent confidence interval:
 1.713251 1.786749
sample estimates:
mean of x
 1.75
```

**Analisar olhando o
p-valor.**

**Se for menor que α ,
rejeita-se H_0 !**

Como $p\text{-valor} < 0,05 = \alpha$, rejeita-se a hipótese nula. Logo, não há evidências com base na amostra para afirmar-se que a média populacional das alturas seja igual a 1,62 metros.

Note que $IC(\mu, 95\%) = [1,71; 1,79]$.

Teste t para amostras independentes



- É utilizado quando deseja-se **comparar médias entre dois grupos**;
- A pressuposição para a realização deste teste é que os dados sejam provenientes de populações que sigam distribuições normais;
- Para a realização do teste **é necessário verificar se as variâncias das duas populações são iguais ou diferentes**, pois para cada situação, uma estatística de teste é utilizada.

Teste t para amostras independentes



Exemplo. Utilizando o mesmo banco de dados do exemplo anterior, verifique a hipótese de que a média das alturas das mulheres é inferior a dos homens. Adote $\alpha = 5\%$ para os testes.

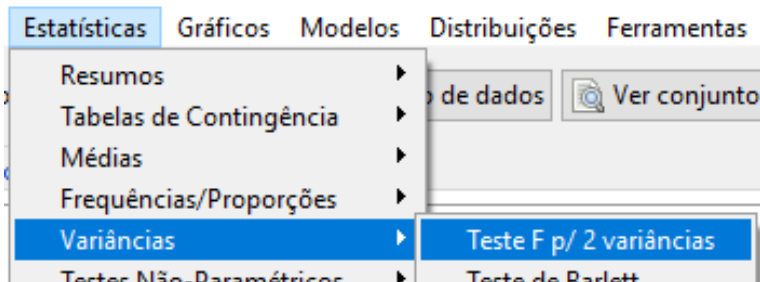
$$H_0 : \mu_{mulheres} = \mu_{homens}$$

$$H_a : \mu_{mulheres} < \mu_{homens}$$

Teste t para amostras independentes

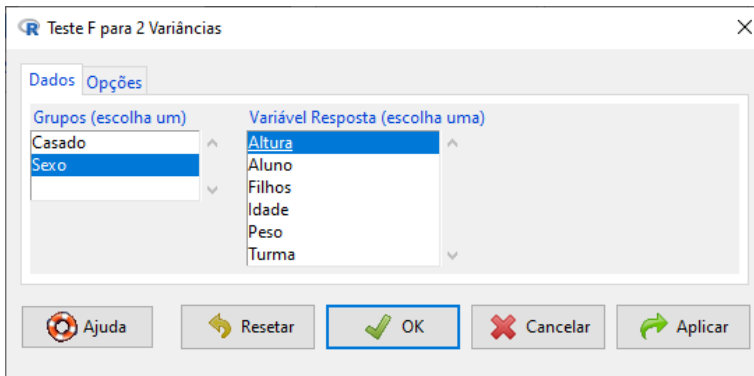
O primeiro passo é descobrir se as variâncias são homogêneas ou não.

Estatísticas > Variâncias > Teste F para 2 variâncias



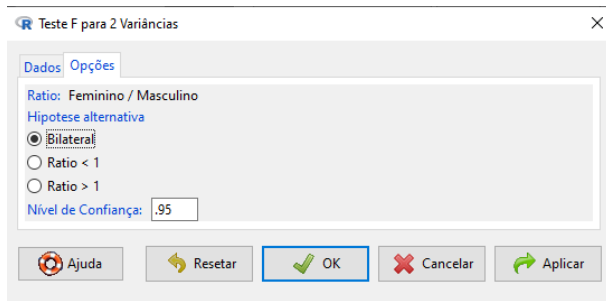
Teste t para amostras independentes

Na aba **Dados**, seleciona-se o grupo (Sexo) e a variável de interesse (Altura).



Teste t para amostras independentes

Na aba **Opções**, selecione a hipótese alternativa.



Com a opção selecionada, as hipóteses são dadas por:

$$H_0 : \sigma_{mulheres}^2 = \sigma_{homens}^2$$

$$H_a : \sigma_{mulheres}^2 \neq \sigma_{homens}^2$$

Teste t para amostras independentes

Output

```
> var.test(Altura ~ Sexo, alternative='two.sided', conf.level=.95, data=aula2)
```

F test to compare two variances

data: Altura by Sexo

F = 0.17356, num df = 7, denom df = 15, **p-value = 0.02608**

alternative hypothesis: true ratio of variances is not equal to 1

95 percent confidence interval:

0.05269852 0.79276323

sample estimates:

ratio of variances

0.1735552

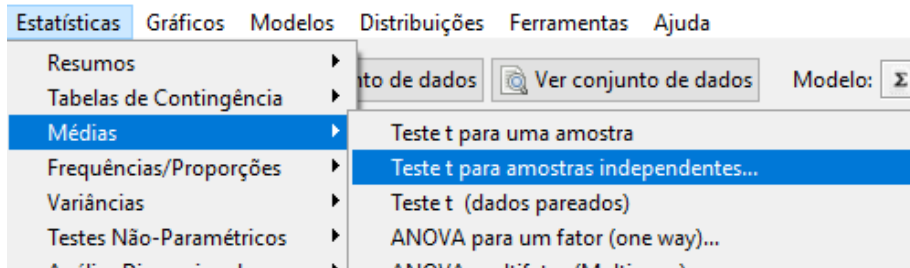
**Analisar olhando o
p-valor.**

**Se for menor que α ,
rejeita-se H_0 !**

Como o $p\text{-valor} < 0,05$, rejeita-se H_0 , e conclui-se, ao nível de 5% de significância, que as variâncias não são homogêneas.

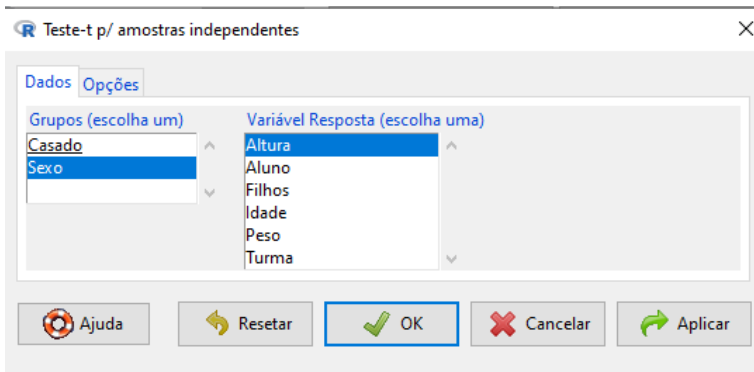
Teste t para amostras independentes

Agora sim, é possível testar se as médias são iguais. Para isso basta seguir os passos:
Estatísticas > Médias > Teste t para amostras independentes



Teste t para amostras independentes

Selecione a variável que especifica os Grupos (Sexo) e a Variável Resposta (Altura), que é a variável com os valores a serem testados.



Teste-t p/ amostras independentes

Dados Opções

Grupos (escolha um)

Casado
Sexo

Variável Resposta (escolha uma)

Altura
Aluno
Filhos
Idade
Peso
Turma

Ajuda Resetar OK Cancelar Aplicar

Teste t para amostras independentes

Teste-t p/ amostras independentes

Dados Opções

Diferença: Feminino - Masculino

Hipotese alternativa	Nível de confiança	Assumir variâncias iguais?
<input type="radio"/> Bilateral	<input type="text" value=".95"/>	<input type="radio"/> Sim
<input checked="" type="radio"/> Diferença < 0		<input checked="" type="radio"/> Não
<input type="radio"/> Diferença > 0		

Ajuda Resetar OK Cancelar Aplicar

Teste t para amostras independentes

Output

```
> t.test(Altura ~ Sexo, alternative = "less", conf.level = .95, var.equal = FALSE, data = aula2)
```

Welch Two Sample t-test

data: Altura by Sexo

t = -4.4067, df = 21.635, p-value = 0.0001156

alternative hypothesis: true difference in means between group Feminino and group Masculino is less than 0

95 percent confidence interval:

-Inf -0.06519838

sample estimates:

mean in group Feminino	mean in group Masculino
1.678750	1.785625

**Analisar olhando o
p-valor.**

**Se for menor que α ,
rejeita-se H_0 !**

Como o p-valor $< 0,05$, rejeita-se H_0 . Assim, pode-se concluir que a altura das mulheres é inferior a altura dos homens, ao nível de 5%.

Teste t (amostras pareadas)



- Uma amostra pareada é aquela na qual tem-se **pares de observações**, exemplos:
 - Medir a pressão arterial de um grupo de pacientes antes e depois de administrar um medicamento.
 - Avaliar o desempenho de estudantes em um teste antes e depois de receberem aulas de reforço.

Teste t (amostras pareadas)

Para testar as médias de amostras pareadas é preciso:

- Calcular a **diferença entre a primeira observação e a segunda observação de cada indivíduo**. Se não houver diferença significativa entre as duas observações para o mesmo indivíduo, a diferença será igual a zero;
- Assumir que a distribuição das diferenças segue uma distribuição normal.

Teste t (amostras pareadas)

Exemplo. Deseja-se avaliar se aulas de reforço de Estatística Básica ministradas para alunos de uma turma da FZEA/USP, foram eficazes para melhorar o rendimento da turma. Registrou-se as notas de cada aluno antes e depois das aulas de reforço. A tabela apresenta os resultados.

Alunos	ANTES	DEPOIS
Laura	2	2
Beatriz	1	2
Julia	8	8
José	0	2
Afonso	10	9
Maria	3	7
Enzo	3	6
Diogo	6	5
Livia	4	3
Cesar	1	3
Antônio	2	5
Felipe	5	7
Kennedy	3	4
Gabriela	0	5
Evelyn	2	7
Ligia	8	8
Yumi	0	5
Rosana	6	9
Valmir	1	5

Teste t (amostras pareadas)

Adotando um nível de $\alpha = 5\%$ de significância, teste se houve um aumento nas notas dos alunos após as aulas de reforço, supondo que a pressuposição de normalidade esteja cumprida.

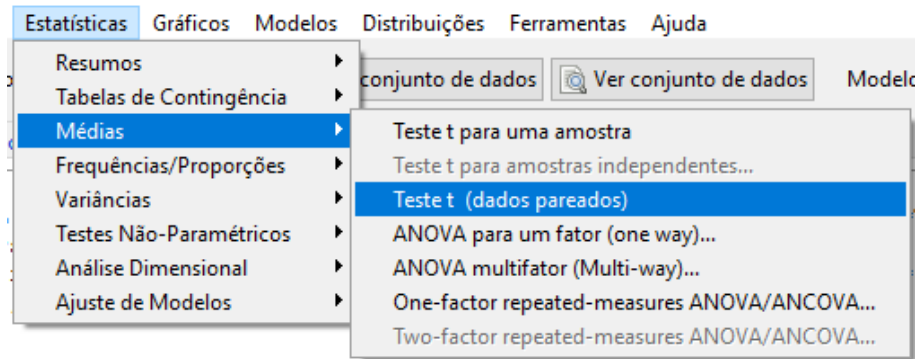
Assim, as hipóteses são dadas por:

$$H_0 : \mu_{depois} = \mu_{antes}$$

$$H_a : \mu_{depois} > \mu_{antes}$$

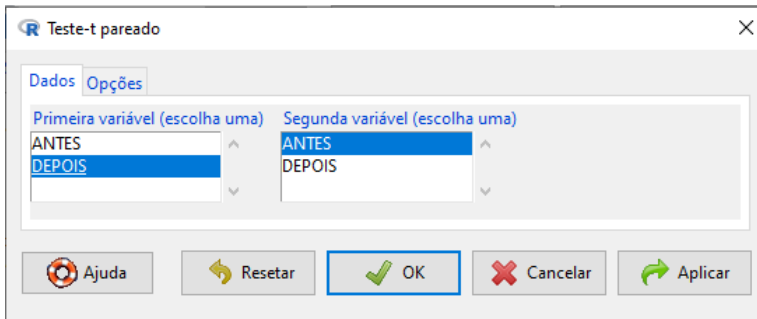
Teste t (amostras pareadas)

No menu, siga o passo a passo: **Estatísticas** > **Médias** > **Teste t (dados pareados)**



Teste t (amostras pareadas)

Na aba **Dados**, selecione as variáveis a serem consideradas, sendo a 1° DEPOIS e a 2° ANTES (pois deseja-se comparar as notas de depois do reforço com antes).



Teste-t pareado

Dados Opções

Primeira variável (escolha uma) Segunda variável (escolha uma)

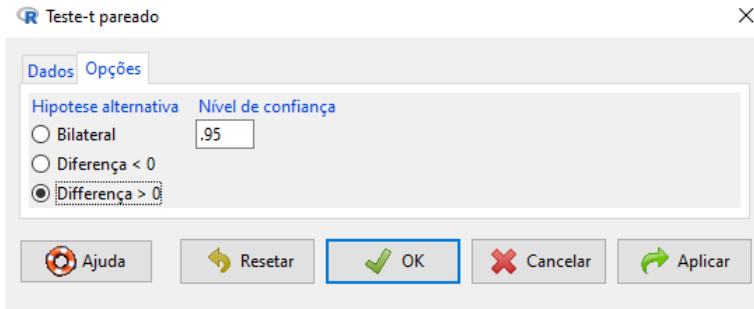
ANTES
DEPOIS

ANTES
DEPOIS

Ajuda Resetar OK Cancelar Aplicar

Teste t (amostras pareadas)

Ainda na aba de **Dados**, seleciona-se o tipo de hipótese alternativa, para este caso, espera-se que a nota de depois seja maior que a nota de antes, ou seja, **espera-se que a diferença seja maior do que zero**.



The image shows a screenshot of the 'Teste-t pareado' (Paired t-test) dialog box in R. The dialog has two tabs: 'Dados' (Data) and 'Opções' (Options). The 'Dados' tab is active. It contains two sections: 'Hipótese alternativa' (Alternative hypothesis) and 'Nível de confiança' (Confidence level). Under 'Hipótese alternativa', there are three radio buttons: 'Bilateral', 'Diferença < 0', and 'Diferença > 0'. The 'Diferença > 0' option is selected. The 'Nível de confiança' is set to '.95'. At the bottom of the dialog, there are five buttons: 'Ajuda' (Help), 'Resetar' (Reset), 'OK', 'Cancelar' (Cancel), and 'Aplicar' (Apply). The 'OK' button is highlighted with a blue border.

Teste t (amostras pareadas)

Output

```
Paired t-test

data: DEPOIS and ANTES
t = 4.0531, df = 18, p-value = 0.000373
alternative hypothesis: true mean difference is greater than 0
95 percent confidence interval:
 1.114223      Inf
sample estimates:
mean difference
 1.947368
```

Analisar olhando o p-valor.

**Se for menor que α ,
rejeita-se H_0 !**

Como $p\text{-valor} < 0,05$, conclui-se, ao nível de 5% de significância, que houve um aumento nas notas dos alunos após as aulas de reforço.



- ▶ Inferência Estatística
- ▶ Teste de normalidade
- ▶ Testes para médias
- ▶ **Teste Qui-quadrado**
- ▶ Testes de Correlação
- ▶ Regressão Linear Simples



Inferência - Teste Qui-quadrado



Um teste qui-quadrado pode ser utilizado para verificar se há dependência entre duas variáveis categóricas.

Exemplo: Com o banco de dados "TURMAGIA2005", verifique se existe uma dependência entre o sexo do estudante e o fato dele ser casado ou não. (Use $\alpha = 5\%$).

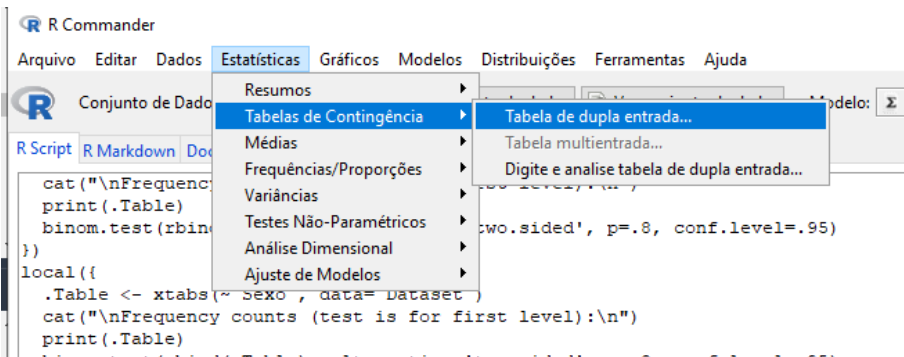
No teste, as hipóteses são dadas por:

H_0 : Estado civil e Sexo são independentes


H_a : Estado civil e Sexo não são independentes

Inferência - Teste Qui-quadrado

No menu, tem-se o caminho: **Estatísticas** > **Tabelas de contingência** > **Tabela de dupla-entrada**



Inferência - Teste Qui-quadrado

 Tabelas de dupla entrada

Dados | Estatísticas

Variável linha (escolha uma)


Casado
Sexo


Variável coluna (escolha uma)


Casado
Sexo


Expressão (subset expression)


<todos casos válidos>

 Ajuda


 Resetar

 OK

 Cancelar

 Aplicar

Inferência - Teste Qui-quadrado

 Tabelas de dupla entrada

Dados Estatísticas

Computar Percentagens






☐ Percentual nas linhas ☐ Percentual nas colunas

☐ Percentagens do total ☒ Sem percentual

Testes de Hipótese

☒ Teste de independência de Qui-Quadrado ☐ Componentes da estatística do Qui-quadrado

☐ Apresente frequências esperadas ☐ Teste exato de Fisher

 Ajuda  Resetar  OK  Cancelar  Aplicar

Inferência - Teste Qui-quadrado

```
Casado Feminino Masculino  
Não      50      31  
Sim      38      71
```

```
Pearson's Chi-squared test
```

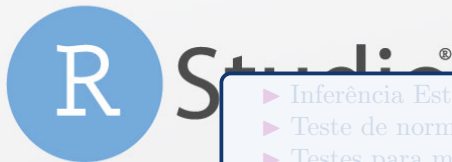
```
data: .Table
```

```
X-squared = 13.489, df = 1, p-value = 0.0002399
```

**Analisar olhando o
p-valor.**

**Se for menor que α ,
rejeita-se H_0 !**

- Como $p\text{-valor} < 0,05$, há evidências de que os fatores estado civil e sexo são dependentes.



- ▶ Inferência Estatística
- ▶ Teste de normalidade
- ▶ Testes para médias
- ▶ Teste Qui-quadrado
- ▶ Testes de Correlação
- ▶ Regressão Linear Simples



Esta medida é usada para medir o grau de associação entre duas variáveis quantitativas.

De forma específica, a correlação de **Pearson** (**r**) mede o grau de correlação **LINEAR** entre duas variáveis quantitativas.

Exemplo: Deseja-se testar se há uma correlação entre peso e altura dos alunos presentes no banco "AULA2".

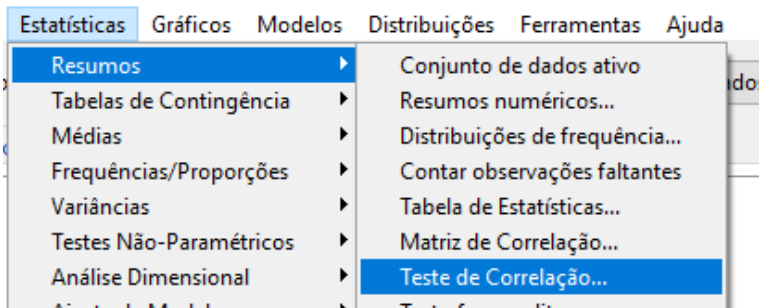
No teste, as hipóteses são dadas por:

H_0 : Não existe correlação entre peso e altura dos alunos

H_a : Existe correlação entre peso e altura dos alunos

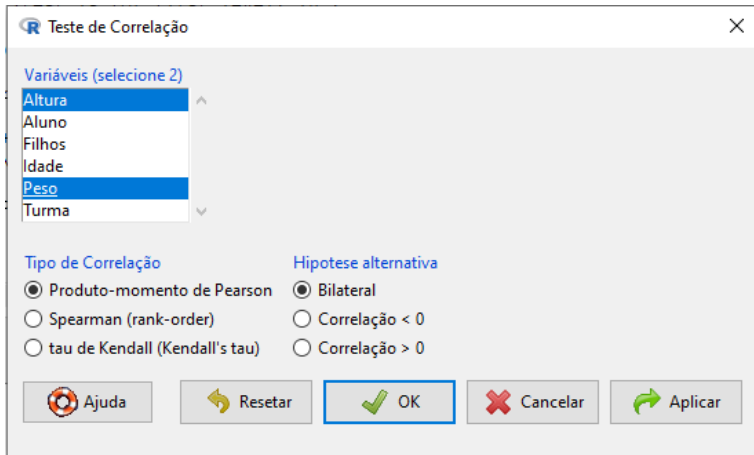
Inferência - Correlação de Pearson

No menu, tem-se que: **Estatísticas** > **Resumos** > **Teste de correlação**



Inferência - Correlação de Pearson

Na janela que abrir, basta selecionar as duas variáveis (com a tecla CTRL seleciona-se a segunda).



Teste de Correlação

Variáveis (selecione 2)

- Altura
- Aluno
- Filhos
- Idade
- Peso
- Turma

Tipo de Correlação

- ☒ Produto-momento de Pearson
- ☐ Spearman (rank-order)
- ☐ tau de Kendall (Kendall's tau)

Hipotese alternativa

- ☒ Bilateral
- ☐ Correlação < 0
- ☐ Correlação > 0

Ajuda Resetar **OK** Cancelar Aplicar

Inferência - Correlação de Pearson

```
Output

> with(Dataset, cor.test(Altura, Peso, alternative="two.sided", method="pearson"))

Pearson's product-moment correlation

data:  Altura and Peso
t = 3.2724, df = 22, p-value = 0.003483
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.2194284 0.7926251
sample estimates:
      cor 
0.5721778
```

**Analisar olhando o
p-valor.**

**Se for menor que α ,
rejeita-se H_0 !**

$r = 0,5721778$ e $p\text{-valor} < 0,05$, o que significa que existe uma correlação linear positiva e significativa entre as duas variáveis.



- ▶ Inferência Estatística
- ▶ Teste de normalidade
- ▶ Testes para médias
- ▶ Teste Qui-quadrado
- ▶ Testes de Correlação
- ▶ Regressão Linear Simples



Regressão Linear Simples



O modelo de regressão linear simples usa apenas uma variável independente (X) para explicar a variável dependente (Y).

Regressão Linear Simples



Exemplo. Deseja-se estabelecer um modelo de regressão linear simples para o banco de dados denominado por "**NOTAS**", o qual contém o número de horas de estudo para uma prova (X) e a nota obtida na referida avaliação (Y).

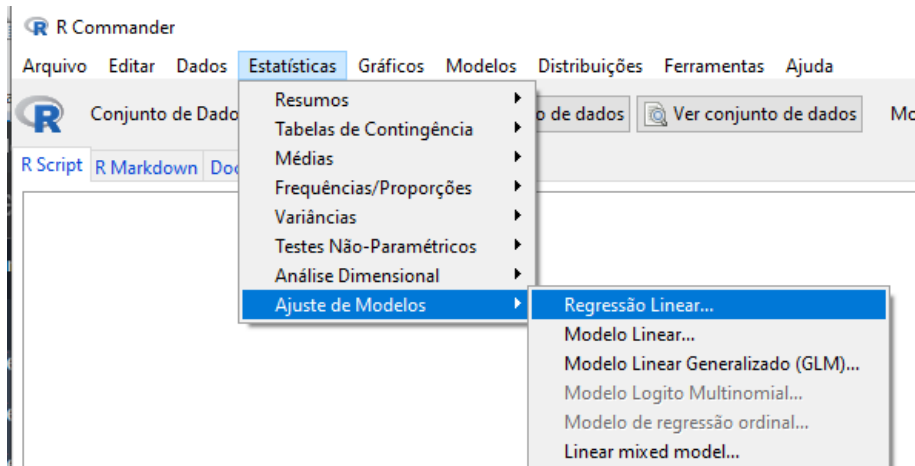
No teste, a principal hipótese de interesse, é testar se o coeficiente angular (β) de uma reta de regressão ($Y = \alpha + \beta X + \epsilon$) pode ser nulo ou não.

$$H_0: \beta = 0$$

$$H_a: \beta \neq 0$$

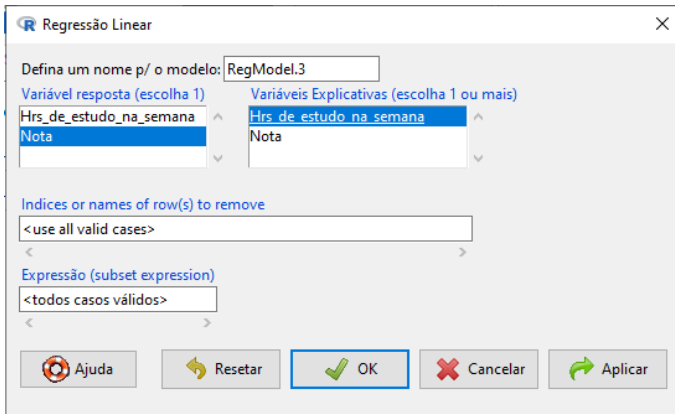
Regressão Linear Simples

No menu, tem-se que **Estatísticas** > **Ajuste de Modelos** > **Regressão Linear...**



Regressão Linear Simples

Selecionar a variável resposta (dependente Y) e a variável explicativa (independente X):



Regressão Linear

Defina um nome p/ o modelo: RegModel.3

Variável resposta (escolha 1)

Hrs_de_estudo_na_semana

Nota

Variáveis Explicativas (escolha 1 ou mais)

Hrs_de_estudo_na_semana

Nota

Indices or names of row(s) to remove

<use all valid cases>

Expressão (subset expression)

<todos casos válidos>

Ajuda

Resetar

OK

Cancelar

Aplicar

Regressão Linear Simples

Output

Call:

```
lm(formula = Nota ~ Hrs_de_estudo_na_semana, data = Dataset)
```

Residuals:

Min	1Q	Median	3Q	Max
-0.98775	-0.06083	-0.02429	-0.00602	0.94831

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-0.03052	0.09389	-0.325	0.747
Hrs_de_estudo_na_semana	2.01827	0.02536	79.597	<2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

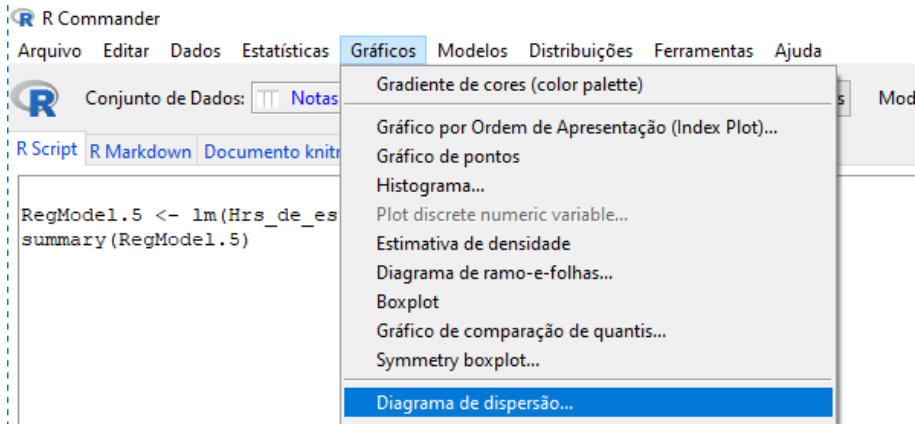
Residual standard error: 0.2598 on 43 degrees of freedom

Multiple R-squared: 0.9933, Adjusted R-squared: 0.9931

F-statistic: 6336 on 1 and 43 DF, p-value: < 2.2e-16

Regressão Linear Simples

Para criar o gráfico da reta de regressão, basta ir em **Gráficos > Diagrama de dispersão**



Regressão Linear Simples

Selecione as variáveis:

Gráfico de Dispersão

Dados Opções

variável-x (escolha uma)

Hrs_de_estudo_na_semana

Nota

variável-y (escolha uma)

Hrs_de_estudo_na_semana

Nota

Gráfico por grupos...

Expressão (subset expression)

<todos casos válidos>

Regressão Linear Simples

E em opções, selecione **Linha de quadrados mínimos**:

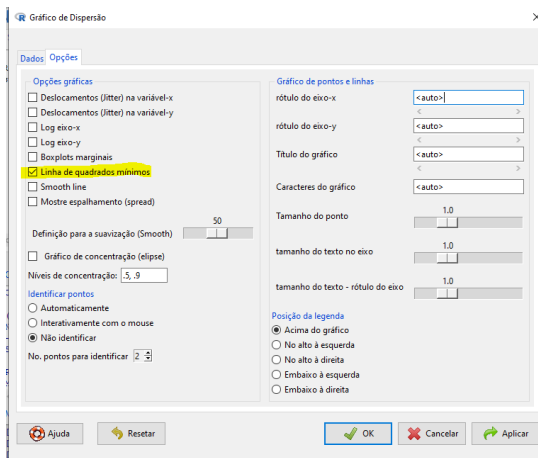


Gráfico de Dispersão

Dados Opções

Opções gráficas

- ☐ Deslocamentos (jitter) na variável-x
- ☐ Deslocamentos (jitter) na variável-y
- ☐ Log eixo-x
- ☐ Log eixo-y
- ☐ Boxplots marginais
- ☒ **Linha de quadrados mínimos**
- ☐ Smooth line
- ☐ Mostre espalhamento (spread)

Definição para a suavização (Smooth) 50

☐ Gráfico de concentração (elipse)

Níveis de concentração: -5, .9

Identificar pontos

- ☐ Automaticamente
- ☐ Interativamente com o mouse
- ☒ Não identificar

No. pontos para identificar 2

Gráfico de pontos e linhas

rótulo do eixo-x <auto>

rótulo do eixo-y <auto>

Título do gráfico <auto>

Caracteres do gráfico <auto>

Tamanho do ponto 1.0

tamanho do texto no eixo 1.0

tamanho do texto - rótulo do eixo 1.0

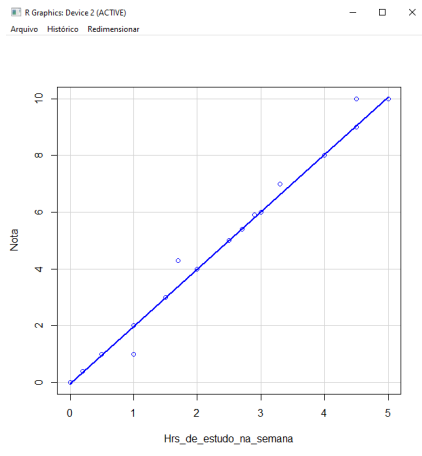
Posição da legenda

- ☒ Acima do gráfico
- ☐ No alto à esquerda
- ☐ No alto à direita
- ☐ Embaixo à esquerda
- ☐ Embaixo à direita

Ajuda Resetar OK Cancelar Aplicar

Regressão Linear Simples

E assim, tem-se:



Referências e links úteis



- Descrição do pacote Rcmdr
- Desvendando a Estatística com o R Commander
- Dicas
- Getting Started With the R Commander
- Graphical Exploration
- Importando dados com R Commander
- O pacote Rcmdr
- R-Studio Vs. Rcmdr