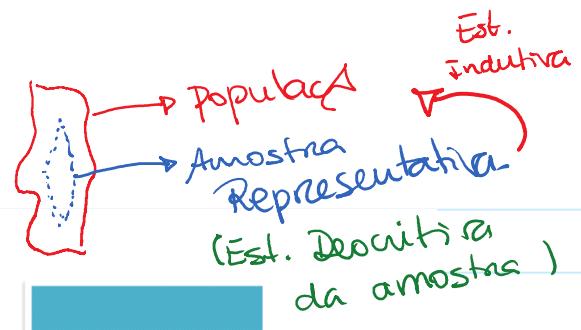


Estatística Descritiva vs Indutiva

28 de abril de 2021 15:17



Estatística Descritiva

- Recolhe
- Classifica
- Organiza
- Apresenta
- Resume



Amostra

Proporcionando índices ou medidas estatísticas

Estatística Indutiva

- Estima
- Prediz
- Compara
- Decide



Métodos Estatísticos

- Estimação pontual
- Estimação intervalar
- Testes de hipóteses

Conceição Rocha / Estatística (Aplicada)

2020/2021

Nota:

Se for conhecida a lei, já conhecemos os parâmetros (média, variância, desvio-padrão populacional)
mas

e se não conhecermos?

→ Como prever?

Pelo estudo da distribuição amostral!

Amostra
Est. Descritiva



Média amostral
Variância amostral
Proporção amostral
(...)

População
Est. Indutiva



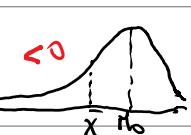
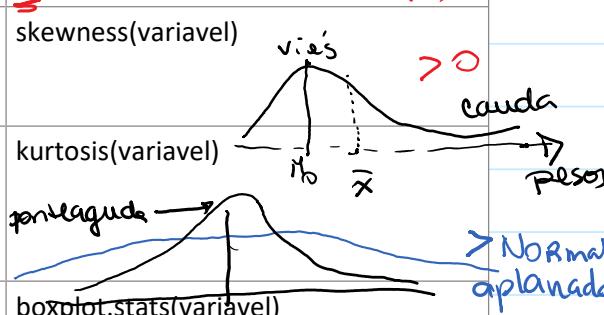
estimacão pontual
estimacão intervalar
Testes

Variáveis

21 de abril de 2021 23:04

Caracterização das Variáveis e tratamento adequado

Variáveis	Tipo	Valores que a variável aleatória X assume:	Exemplos	Gráfico	Estatísticas
Qualitativas	Nominais	Conjunto de nomes (sem ordem)	Cor de olhos; Concelho; Género; Situação profissional; (...)	Circular; Barras;	Frequências Absolutas ou relativas; Moda;
	Ordinais	Características mensuráveis (com ordem)	Concordância - Escalas de Likert; Preferências; Intensidade da dor; (...)	Barras; Diagrama de extremos e quartis;	Frequências Absolutas ou relativas; Moda; Mediana; Quantis; Amplitude amostral; Amplitude Interquartil;
Quantitativas	Discretas	Conjunto numérico finito ou infinito numerável	Número de pessoas; Quantidades; (...)	Barras; Diagrama de extremos e quartis;	Frequências Absolutas ou relativas; Moda; Média, Mediana; Quantis; Amplitude amostral; Amplitude Interquartil; Variância, Dispersão; Assimetria; Achatamento; Outliers
	Contínuas	Conjunto numérico dentro de um determinado intervalo ou conjunto infinito não numerável	Tempo; Peso; Comprimento; (...)	Histogramas; Diagrama de extremos e quartis;	Frequências Absolutas ou relativas; Moda; Média, Mediana; Amplitude amostral; Amplitude Interquartil; Quartis; Variância, Dispersão; Assimetria; Achatamento; Outliers

Medidas		Comando R
Contagens	Frequências Absolutas (unidades); Frequências Relativas (percentagens);	table(variavel) prop.table(table(variavel))
Localização Central:	Moda - mais frequente; Média - valor em torno do qual a distribuição se encontra (menos resistente) Mediana - por ordem crescente é valor do meio localiza 50% (mais resistente à influência de outliers);	table(variavel) mean(variavel) median(variavel) Ou quantile(variavel, 0.5)
Localização Não central:	Quantis - dividem a amostra em n partes iguais; Quartis - 4 partes iguais a 25%: Ex.: O 2º quartil é a mediana, o terceiro acumula 75% Decis - 10 partes iguais a 10%; Percentis - 100 partes iguais a 1%;	summary(variavel) (quartis e média) quantile(variavel) (quartis) quantile(variavel, 0.85) (Percentil 85)
Dispersão:	Amplitude amostral: Max-Min (100%); Amplitude Interquartil: Q3-Q1 (50%); Variância; Desvio-padrão; ...	max(variavel) -min(variavel) var(variavel) sd(variavel) $D.P = \sqrt{v(x)}$
Assimetria	Simetria estuda Viés e caudas; Positiva: média>moda Negativa: média<moda	skewness(variavel) 
Curtose	Compara o achatamento provocado pela variância Mesocúrtica: igual à Normal Leptocúrtica: ponteaguda Platicúrtica: mais plana	kurtosis(variavel) 
Outliers	Moderados; Severos;	boxplot.stats(variavel)
Gráficos:		
Circular	Variáveis qualitativas nominais (grupos!!!)	pie(table(variavel))
Barras	Variáveis qualitativas nominais Variáveis qualitativas ordinais Variáveis quantitativas discretas	barplot(table(variavel)) boxplot(variavel)
histograma	Variáveis quantitativas contínuas	hist(variavel) boxplot(variavel)
Comparação entre grupos	variáveis numéricas por grupos	boxplot(variavel_num ~ var_grupos) tapply(var_num, var_grupos, summary) table(genero, idade)

Cruzamentos - comparar grupos e distribuições

22 de abril de 2021 11:15

grupos (Mulher, Homem)
obedece / não obedece a determinadas
características

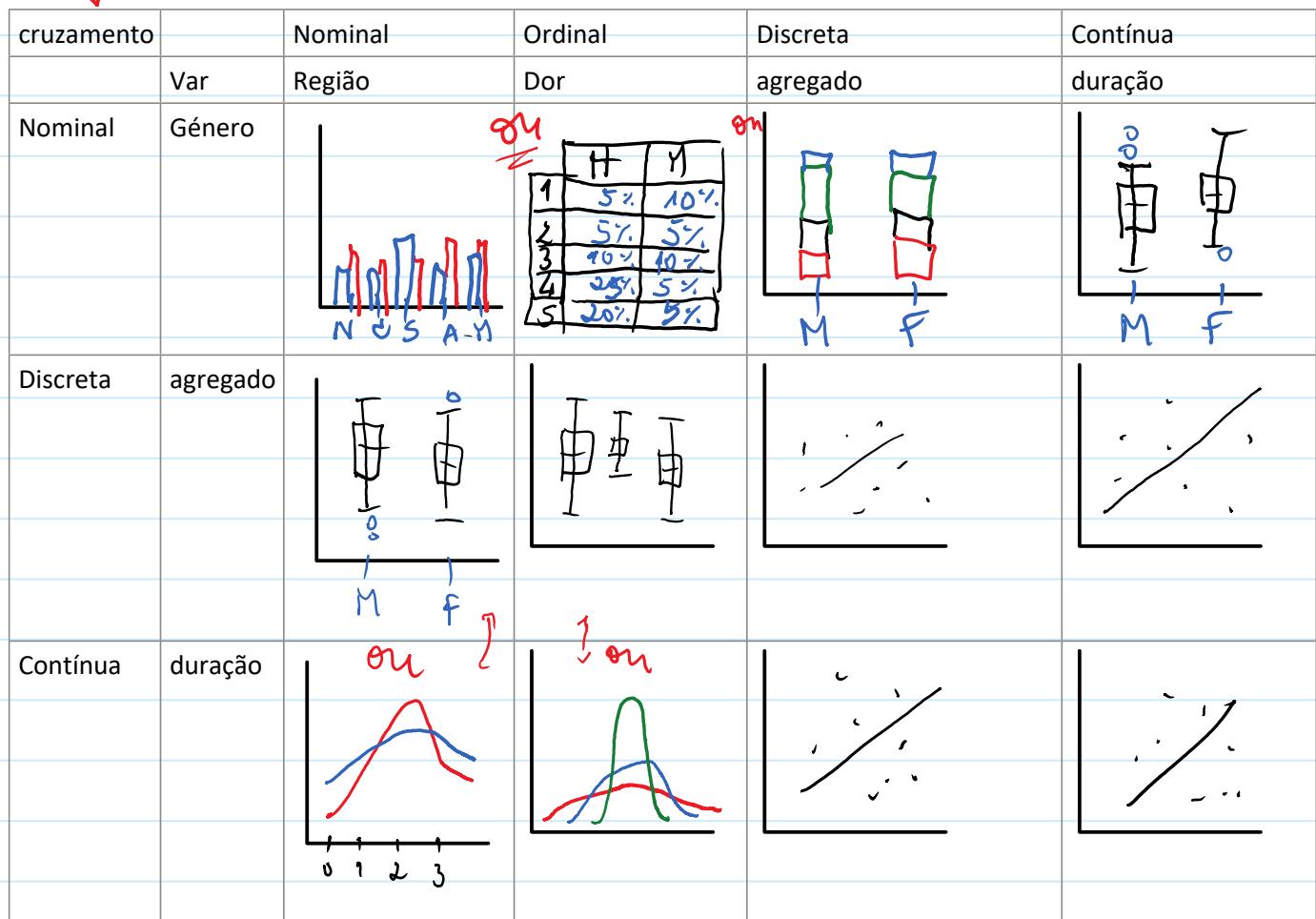
Var.	Nominal	Ordinal	Discreta	Contínua
Nominal	Tabela Dupla entrada Gráficos empilhados	Tabela Dupla entrada Gráficos empilhados		
Ordinal	Tabela Dupla entrada Boxplot Múltiplo	Tabela Dupla entrada Boxplot Múltiplo		
Discreta	Boxplot Múltiplo	Boxplot Múltiplo	Nuvem de Pontos	Nuvem de Pontos
Contínua	Boxplot Múltiplo	Boxplot Múltiplo	Nuvem de Pontos	Nuvem de Pontos

→
não filhos
Salário →

Quantitativo
↓

VS

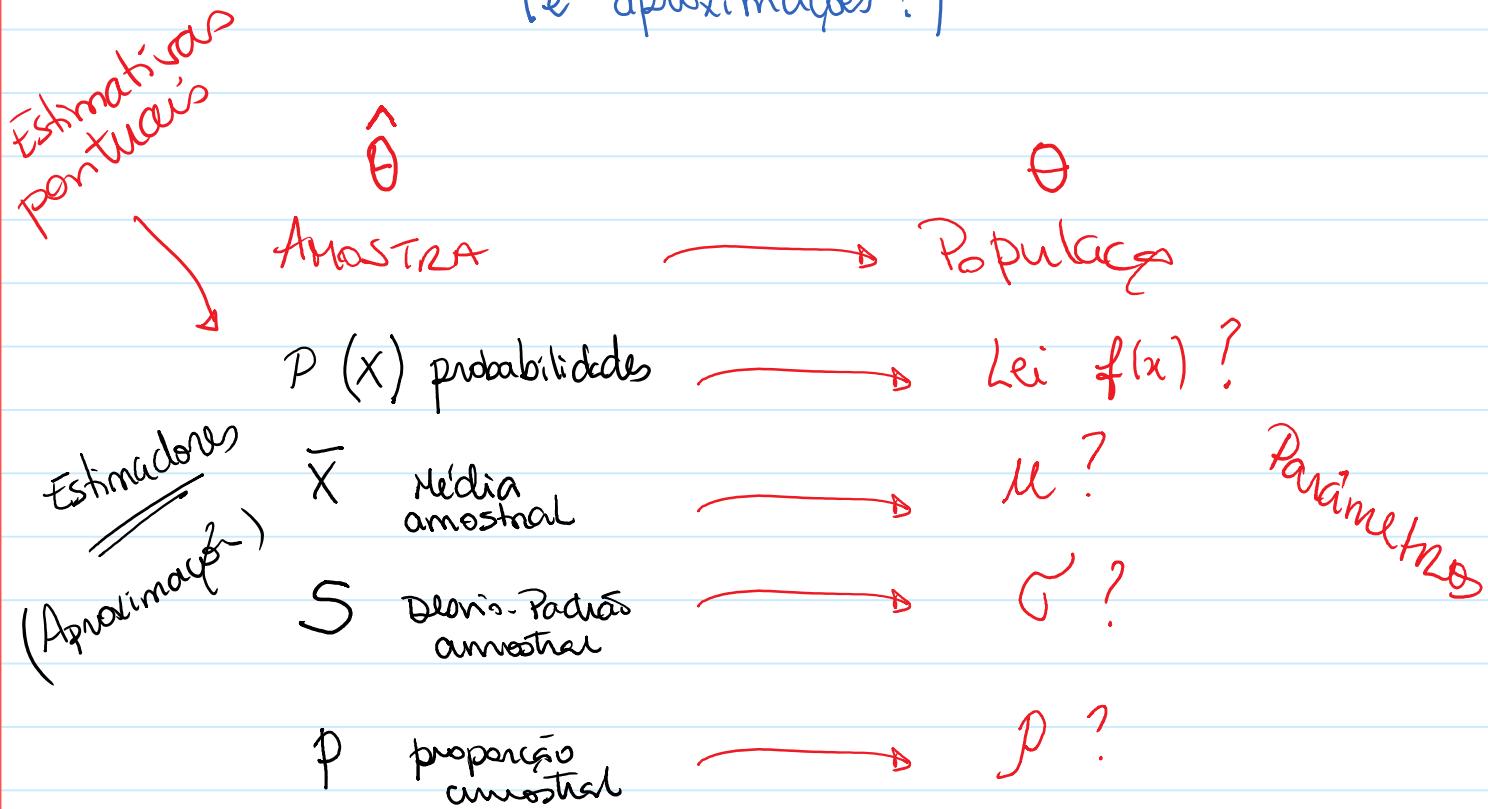
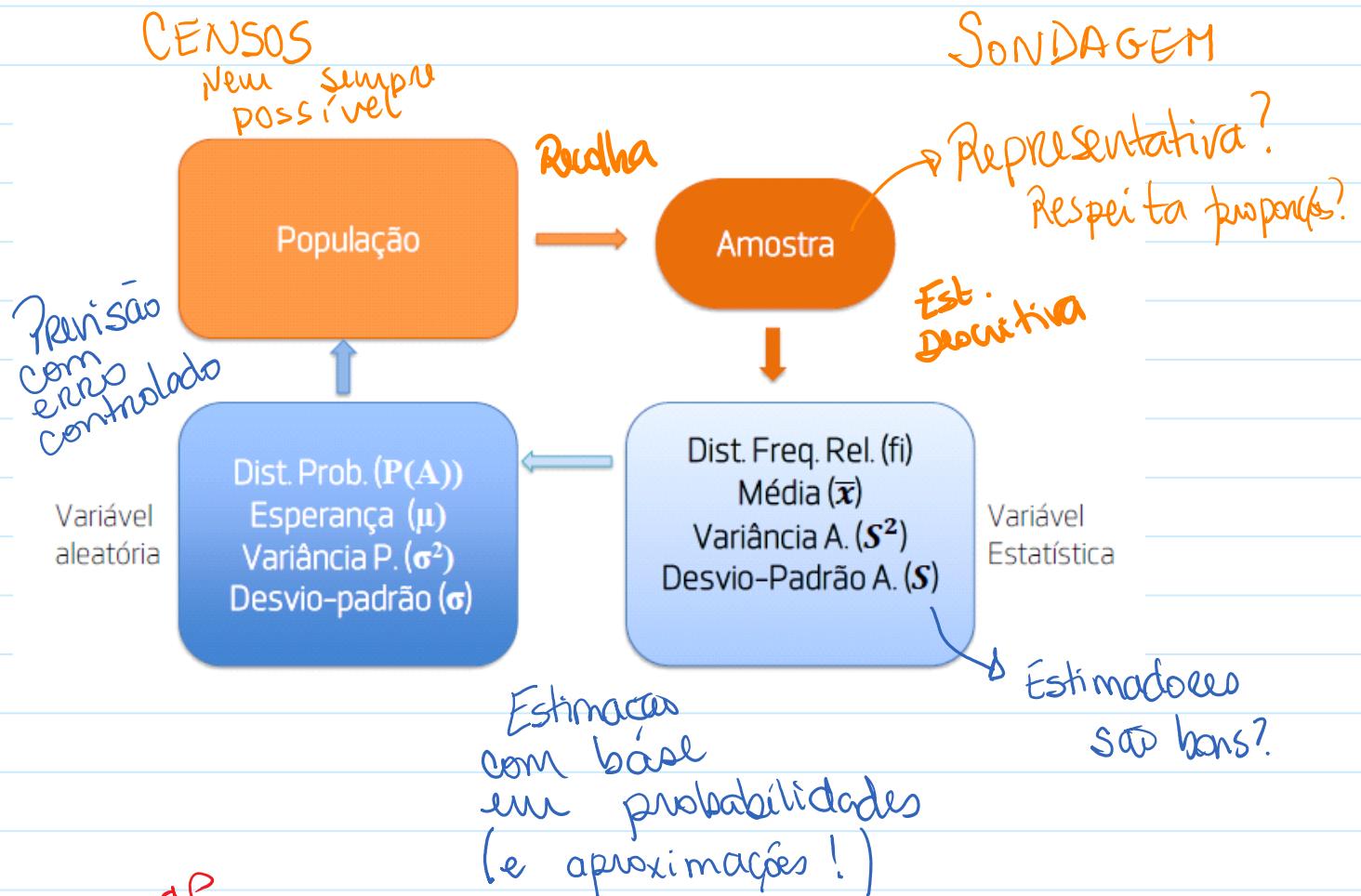
Qualitativas (palavras)
Grupos!



Processo de Estimação

29 de abril de 2021 10:16

? Como Prever ?



Estimação pontual e intervalar

12 de maio de 2021 18:10

Para estimar o valor da média populacional μ toma-se a média amostral (\bar{x})

Mas $\bar{x} \neq \mu$ na maioria das amostras!

A Estimação pontual tem um erro elevado.



Estimação intervalar:

- Toma um estimador do parâmetro θ
p. ex.: \bar{x} é estimador de μ
- estuda a sua distribuição
- Considera um nível de significância (α)
- Define os limites do intervalo (quantis).

do tipo:

$$[t_1, t_2] : P(t_1 < \theta < t_2) = 1 - \alpha$$

grau conf.

Como definir t_1, t_2 ?

Exemplo:

Conhecidos 1º Q e 3º Q sabemos que

$$P(\theta \in [1^\circ Q ; 3^\circ Q]) = 50\% \quad (N \cdot S = 50\%)$$

I. Conf. insatisfatório.

Objetivo: Diminuir o N. Sign. α
Diminuir o Intervalo

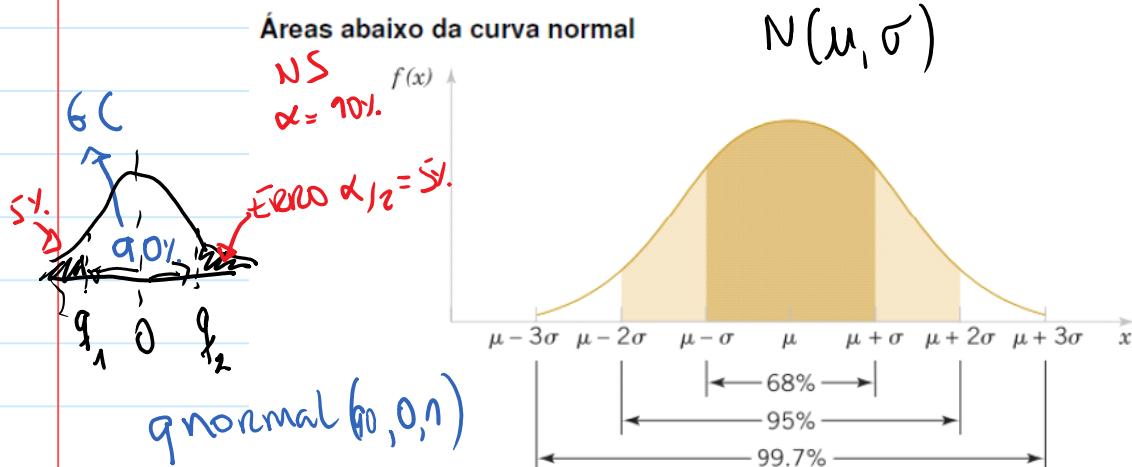
↑
erro
mt grande

Propriedades da Lei Normal

29 de abril de 2021 11:32

Como baixar o erro? Através da lei Normal

A lei Normal é simétrica e muito concentrada em torno da média, dai podemos retirar os seguintes Intervais de Confiança:



$$Z = \frac{\bar{X} - \mu}{\sigma}$$

$Z \sim N(0, 1)$

Normalização da variável: $\frac{\bar{X} - \mu}{\sigma} \sim N(0, 1)$

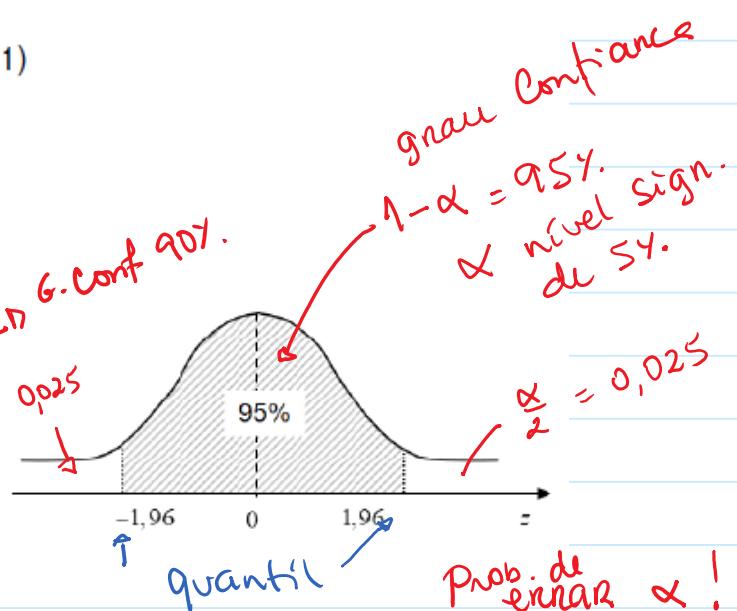
Distribuição Normal Padrão: $Z \sim N(0, 1)$

$$\text{qnorm}\left(\frac{\alpha}{2}, 0, 1\right)$$

$\alpha = -1,645$
 $\alpha = -1,95$
 $\alpha = -2,5758$

Intervalo	Probabilidade
$\mu \pm \sigma$	0,6826
$\mu \pm 2\sigma$	0,9544
$\mu \pm 3\sigma$	0,9973
$\mu \pm 0,6745\sigma$	0,5000
$\mu \pm 1,6450\sigma$	0,9000
$\mu \pm 1,9600\sigma$	0,9500
$\mu \pm 2,5758\sigma$	0,9900

valor de z (quantil)
grau confiança



Verifica para qualquer $N(\mu, \sigma)$

Intervalo de Confiança	Nível de Confiança	Nível de significância
$[\mu - 1,645\sigma, \mu + 1,645\sigma[$	90%	10%
$[\mu - 1,96\sigma, \mu + 1,96\sigma[$	95%	5%
$[\mu - 2,5758\sigma, \mu + 2,5758\sigma[$	99%	1%

Resumo intervalos

I.C.R.

Estimar

média

μ

Média

pop.

→ Conhecido $Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$ → Z-test

$\sigma_{\text{pop.}}$

$N(0,1)$

Simétrica

`z.test(var_X, sigma.x=sd, conf.level = 0.90)`

→ Desconhecido $T = \frac{\bar{x} - \mu}{s/\sqrt{n}} \sim t(n-1)$ → t-test

$s_{\text{pop.}}$

Simétrica

`t.test(var_X, conf.level = 0.95)`

Variância
pop.

$\chi^2(n-1)$

$\rightarrow \sigma^2 \in]\frac{(n-1)s^2}{9x^2(1-\alpha)}, \frac{(n-1)s^2}{9x^2(\alpha)}[$

\uparrow
não
simétrica!
 α_1 e α_2 são diferentes!

não tem
teste p/
I. Conf.
(à mão!)

TLC

Proporção
pop. $\rightarrow X \sim B(n, p) \sim N(p, \sqrt{pq}) \rightarrow \text{prop.test}$

prop Sucesso
amostra

= $\frac{n^o \text{ Sucessos}}{n \text{ amostra}}$

Ex:

`prop.test(x=47, n=83, conf.level=0.96)`

47 sucedeu em 83

calcular com `table(var_qual)`

entre grupos (1 var qualitativa)

$\mu_1 = \mu_2 ?$
ou $\mu_1 - \mu_2 = 0$

→ σ conhecido

→ σ desconhecido

→ Z-test (G_1, G_2)

→ T-test (G_1, G_2)

$\sigma_1^2 = \sigma_2^2 ?$
ou $\frac{\sigma_1^2}{\sigma_2^2} = 1$

→ var. test (G_1, G_2)

$p_1 = p_2 ?$

→ prop. test (G_1, G_2)

Resumo testes

13 de maio de 2021 10:51

rejeitar H_0 : p-value < α

(conce-se o risco de rejeitar porque o erro tipo I mat. é mais pequeno do que o N.S. α)

Existem evidências estatísticas de que se deve rejeitar H_0
Não existem a um N.S. de ... %?

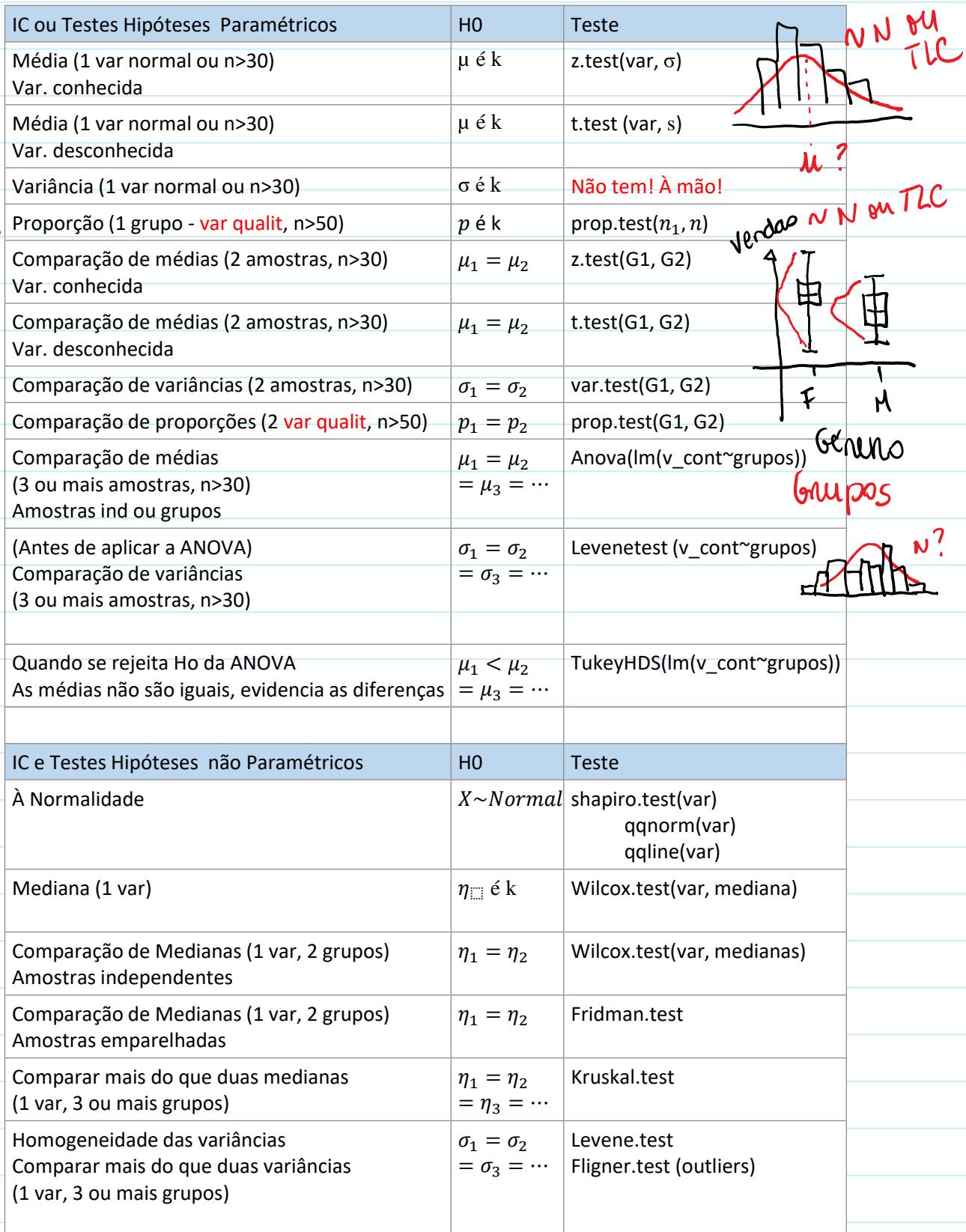
IC ou Testes Hipóteses Paramétricos	H_0	Teste
Média (1 var normal ou $n > 30$) Var. conhecida	$\mu \text{ é } k$	<code>z.test(var, sigma)</code>
Média (1 var normal ou $n > 30$) Var. desconhecida	$\mu \text{ é } k$	<code>t.test(var, s)</code>
Variância (1 var normal ou $n > 30$)	$\sigma \text{ é } k$	Não tem! À mão!
Proporção (1 grupo - var qualit, $n > 50$)	$p \text{ é } k$	<code>prop.test(n1, n)</code>
Comparação de médias (2 amostras, $n > 30$) Var. conhecida	$\mu_1 = \mu_2$	<code>z.test(G1, G2)</code>
Comparação de médias (2 amostras, $n > 30$) Var. desconhecida	$\mu_1 = \mu_2$	<code>t.test(G1, G2)</code>
Comparação de variâncias (2 amostras, $n > 30$)	$\sigma_1 = \sigma_2$	<code>var.test(G1, G2)</code>
Comparação de proporções (2 var qualit, $n > 50$)	$p_1 = p_2$	<code>prop.test(G1, G2)</code>
Comparação de médias (3 ou mais amostras, $n > 30$) Amostras ind ou grupos	$\mu_1 = \mu_2 = \mu_3 = \dots$	Anova(<code>lm(v_cont ~ grupos)</code>)
(Antes de aplicar a ANOVA) Comparação de variâncias (3 ou mais amostras, $n > 30$)	$\sigma_1 = \sigma_2 = \sigma_3 = \dots$	<code>Levenetest(v_cont ~ grupos)</code>
Quando se rejeita H_0 da ANOVA As médias não são iguais, evidencia as diferenças	$\mu_1 < \mu_2 = \mu_3 = \dots$	<code>TukeyHDS(lm(v_cont ~ grupos))</code>
IC e Testes Hipóteses não Paramétricos	H_0	Teste
À Normalidade	$X \sim \text{Normal}$	<code>shapiro.test(var)</code> <code>qqnorm(var)</code> <code>qqline(var)</code>
Mediana (1 var)	$\eta \text{ é } k$	<code>Wilcox.test(var, mediana)</code>
Comparação de Medianas (1 var, 2 grupos) Amostras independentes	$\eta_1 = \eta_2$	<code>Wilcox.test(var, medianas)</code>
Comparação de Medianas (1 var, 2 grupos) Amostras emparelhadas	$\eta_1 = \eta_2$	<code>Friedman.test</code>
Comparar mais do que duas medianas (1 var, 3 ou mais grupos)	$\eta_1 = \eta_2 = \eta_3 = \dots$	<code>Kruskal.test</code>
Homogeneidade das variâncias Comparar mais do que duas variâncias (1 var, 3 ou mais grupos)	$\sigma_1 = \sigma_2 = \sigma_3 = \dots$	<code>Levene.test</code> <code>Fligner.test(outliers)</code>

↑ Var
Normal
ou
TLC

↓ Var
N
ou
TLC

↓ Var qual
+ do que
2 grupos

NÃO
SEGUE
Normal
ou
 $n < 30$.
Aproximar
a média
à mediana



Esquema Testes

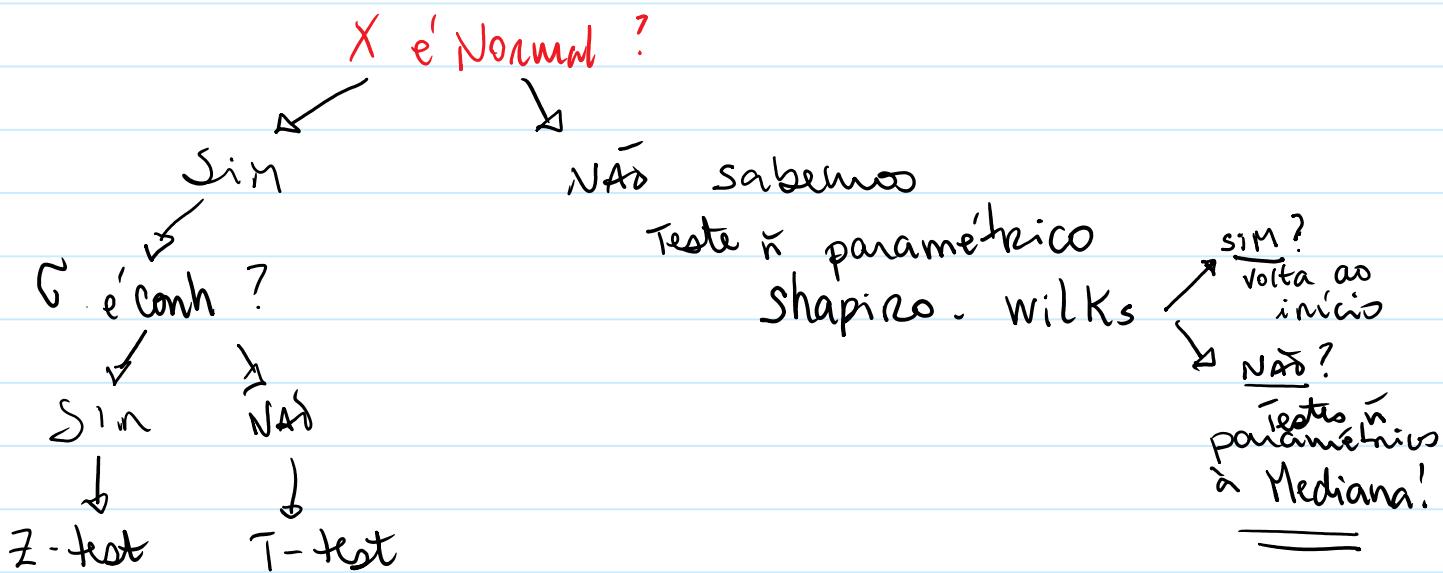
4 de junho de 2021 00:10

Passo 1: $\theta = ?$

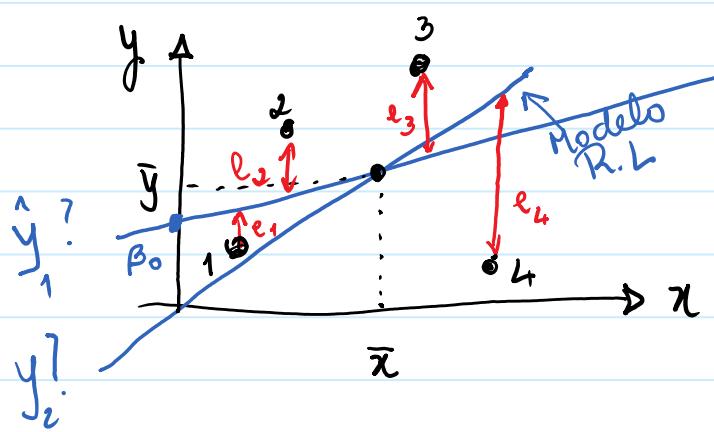
O que pretendemos estimar? Média? Variâncias? Coeficientes? Proporções?

Passo 2: Os grupos/vars? 1, 2 ou mais?

Passo 3: Normalidade? $> 30?$



Estudo da relação (associação) entre duas variáveis quantitativas



A MOSTRA (x, y)
→ Nuvem de pontos

Definir o Modelo \hat{y}_1, \hat{y}_2
Ajustar ao melhor \hat{y}

(?) Minimizar os erros

Inferir y

1º) y é a variável dependente

x é a variável independente

2º) $y = \beta_0 + \beta_1 \cdot x + \epsilon$, β_0 e β_1 - coeficientes de Regressão

↳ Eq. Reta de Regressão onde:

β_0 e β_1 são constantes a estimar (significativas?)

β_0 é o intercept (ordenada na origem)

β_1 é o declive (variação de y em torno de x)

ϵ São os erros provocados pelo modelo que devem ser minimizados e controlados

Coeficientes de correlação e determinação

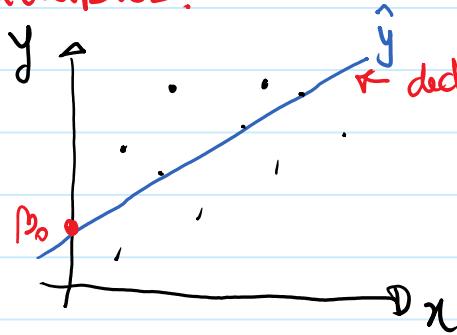
25 de maio de 2021 01:05

→ r - Coeficiente de Correlação (r)

Coeficiente de correlação (de Pearson) – permite avaliar o grau de associação linear entre duas variáveis

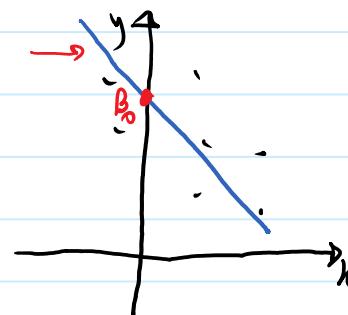
$$r = \frac{s_{xy}}{s_x s_y} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{s_x s_y},$$

Exemplos:



$$Y = \beta_0 + \beta_1 X + \varepsilon$$

Relação positiva



$$Y = \beta_0 - \beta_1 X + \varepsilon$$

Relação negativa

classificam

Coeficiente de correlação	Correlação
$r = 1$	Perfeita positiva
$0,8 \leq r < 1$	Forte positiva
$0,5 \leq r < 0,8$	Moderada positiva
$0,1 \leq r < 0,5$	Fraca positiva
$0 < r < 0,1$	Ínfima positiva
$r = 0$	nula
$-0,1 < r < 0$	Ínfima negativa
$-0,5 < r \leq -0,1$	Fraca negativa
$-0,8 < r \leq -0,5$	Moderada negativa
$-1 < r \leq -0,8$	Forte negativa
$r = -1$	Perfeita negativa

→ r^2 - Coeficiente de determinação: mede a qualidade ajustamento



Coeficiente de determinação – representa a proporção da variável Y que é explicada pela regressão

$$R^2 = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = \frac{\text{variação de } Y \text{ explicada pela regressão}}{\text{Variação total de } Y}, \text{ com } 0 \leq R^2 \leq 1$$

fazer Sempre a interpretação.

→ Nota: r_a^2 - Coeficiente de det. ajustado: é melhor quando há mais do que uma var. independente (R. linear múltipla)

Podemos usar os modelos mas temos que verificar os pressupostos:

Os erros seguem uma Variável Aleatória NORMAL e São independentes

A análise ao nível inferencial requer que:

- Os erros tenham média zero e variância (σ^2) constante (homocedasticidade)
- Os erros são mutuamente independentes (autocorrelação nula)
- Os erros seguem a distribuição normal

Resumo Regressão

25 de maio de 2021 01:21

Em R:

Duas variáveis quantitativas

Modelo Reg. Linear $y = \beta_0 + \beta_1 x + \epsilon$

plot(x, y)

modelos = lm(y ~ x)

reta = abline(modelos)

- Descriptivo

- Nuvem de n pontos (x_i, y_i) no plano → **diagrama de dispersão**
- Reta que passa o mais próximo possível da nuvem de pontos → **reta de regressão**
- Sentido e força da relação linear entre as variáveis X e Y → **coeficiente de correlação**

summary(modelos)

→ c. cor(x, y)

→ c. Det(x^2)

→ c. Det. ajustado(x_a^2)

- Inferencial

- Verificação da existência de relação linear → teste **ANOVA**
- Confirmação da adequação do modelo → estudo dos **resíduos**
- Previsão de valores de Y para valores de X fixos → intervalos de **previsão**

↓
Coeficientes
(ANOVA)

Pressupostos:

Estudo dos Resíduos:

Normalidade: Shapiro.test
 $H_0: \epsilon \sim N(0, \sigma^2)$

Autocorrelação
nula: Durbin-Watson
 $H_0: \rho = 0$ (Não há autocorrelação)

Independência

São independentes

Previsão:

$$\hat{y} = \beta_0 + \beta_1 \cdot x$$

\hookrightarrow Inférfia

(\hat{y} - approximado)

↳ Resolver em ordem a X ou Y.

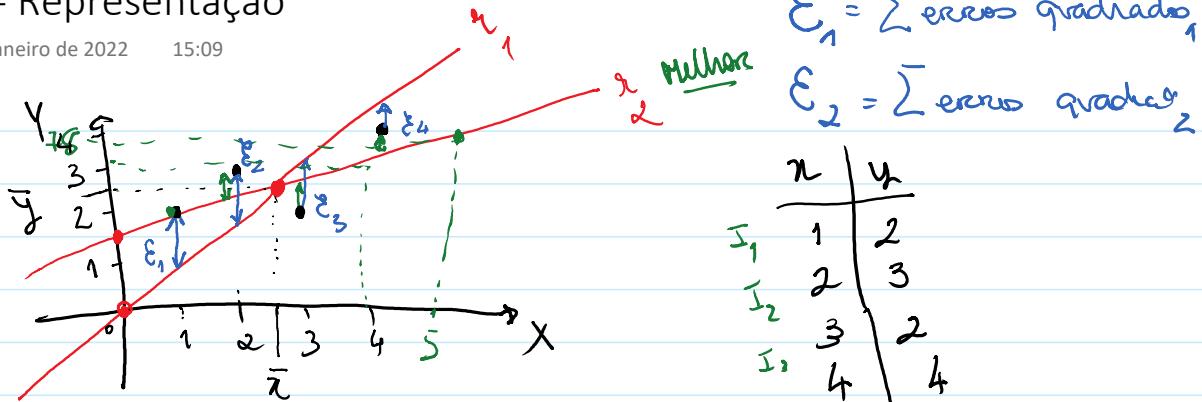
Duas variáveis qualitativas

Teste do Qui - quadrado

Chiqr. test

Ex - Representação

5 de janeiro de 2022 15:09



$$\Sigma \text{ erros quadrados}_1$$

$$\Sigma \text{ erros quadrados}_2$$

x	y
I ₁	2
I ₂	3
I ₃	2
I ₄	4
I ₅	5

$$\bar{x} = 2,5$$

$$\bar{y} = 2,52$$

Eq. reta:

$$y = ax + b \quad , a \neq b \text{ cte}$$

↓
declive

↓ ordenadas na origem

(aproximado) $\hat{Y} = 1,8 + 1,2x$

) No R:
 intercept: 1,8 (ord. na origem)
 x_0 : 1,2 (declive)

$y = 1,8 + 1,2x + \epsilon$

\hat{I}_5 : $x=5$ anos serviço. $\rightarrow y = \text{Nº dias férias}$

$$y = 1,8 + 1,2 \times 5 = 7,8$$

\uparrow
 x_5 \uparrow
 y_5

\hat{I}_6 : 3,5 dias de férias $\rightarrow 3,5 = 1,8 + 1,2 x_6$?

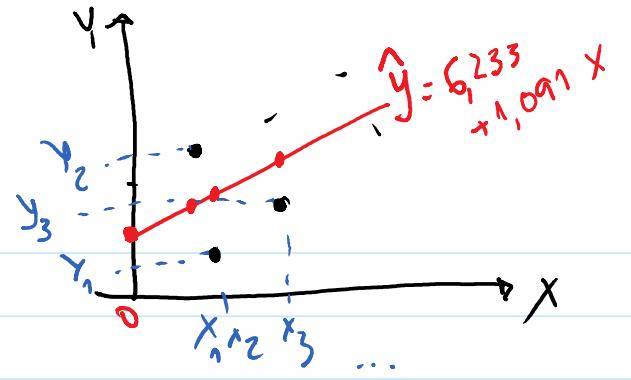
$$x_6 = \frac{3,5 - 1,8}{1,2}$$

Exercicio 1 em R

5 de janeiro de 2022 16:36

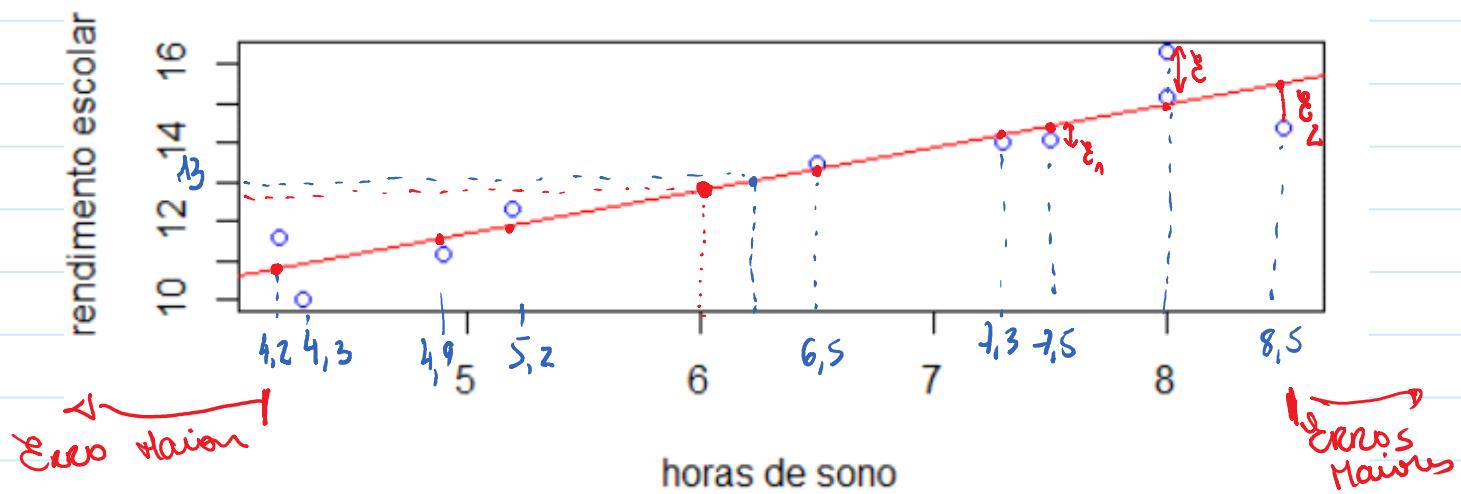
$$Y = \text{intercept} + b * x + \text{erros}$$

$$6.233 + 1.091 * \text{horas de sono}$$



predict (RL)

Rendimento VS Sono



Coefficientes :

De correlação (r) : mede a associação entre X e Y
 $-1 \leq r \leq 1$

De Determinação (r^2) : ajustamento ao modelo

De Regressão : intercept + coeficiente em X

Interpretação Coeficientes

5 de janeiro de 2022 16:55

rl → Modelo Reg. Linear

> summary(rl)

Call:

lm(formula = rend ~ sono)

Residuals:

Min	1Q	Median	3Q	Max
-1.10779	-0.36387	<u>-0.01193</u>	0.35423	1.33779

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	6.233	1.060	5.879	0.000370 ***
sono	1.091	0.160	6.821	0.000135 ***

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

Residual standard error: 0.7916 on 8 degrees of freedom
Multiple R-squared: 0.8533, Adjusted R-squared: 0.8349
F-statistic: 46.53 on 1 and 8 DF, p-value: 0.0001349

R² Corf. Det.

Segundo esta amostra e este modelo, a variância das horas de sono explica 85,33% da var. do rend. escolar.

$$y = \beta_0 + \beta_1 x + \epsilon$$

$$\hat{y} = \beta_0 + \beta_1 x$$

→ Muito próximos de zero (Normal de média 0)?

Modelo: $\text{rend} = 6,233 + 1,091 \times \text{sono}$

1º Se $X=0 \rightarrow \hat{\text{rend}} = 6,233$: Se o nº de horas de sono for zero o rend esperado pelo modelo será de 6,233 valores (o que não parece suportável) Note-se que $0 \notin [4,2 ; 8,5] \rightarrow$ Janela de observações

2º $\rightarrow 1,091$: Por cada hora de sono que, em média, o estudante dorme a mais prevê-se um aumento de 1,091 valores na média escolar.

* → Testes aos coeficientes do modelo:

$$H_0: \beta = 0$$

$$H_1: \beta \neq 0$$

Em ambos $p\text{-value} < 0,05$ rejeita-se a H_0 , i.e., não existem evidências estatísticas de que os coeficientes se anulem logo o modelo é válido.

Interpretação dos coeficientes

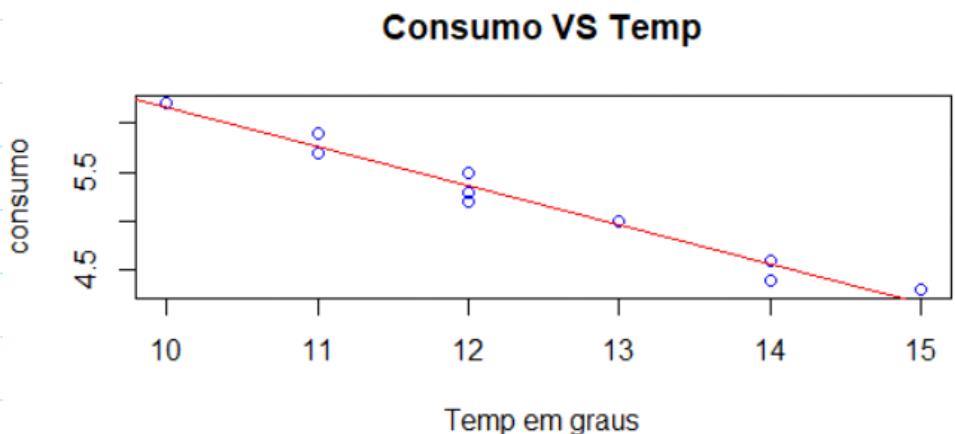
5 de janeiro de 2022 17:29

Call:
lm(formula = cons ~ temp)

Coefficients:

(Intercept)	temp
10.1589	-0.3991

Coeficiente
independ.



Modelo:

$$y = \alpha x + b$$

$$\hat{C}\text{ons} = -0,3991 \times \text{temperatura} + 10,1589$$

Interpretacão:

Se a temperatura se anular, o consumo previsto é de 10,1589 kwh.

Por cada grau que a temperatura suba prevê-se que o consumo a cresça 0,3991 kwh.

cons = 10.1589 - 0.3991 *temperatura

10.1589 - 0.3991 *13.5 #4.77105 kwh

10.1589 - 0.3991 * 42 #-6.6033 kwh

10.1589 - 0.3991 * 16#= 3.7733 kwh se a temperatura for de 16 graus espera-se um consumo de 3.77

Estudo dos erros

24 de maio de 2021 18:10

4.5. Estudo dos Resíduos \Rightarrow Normalidade

Gráficamente parece Normal
Teste Shapiro-Wilk:

$H_0: \epsilon \sim \text{Normal}$

Shapiro-Wilk normality test

data: modelolinear\$residuals

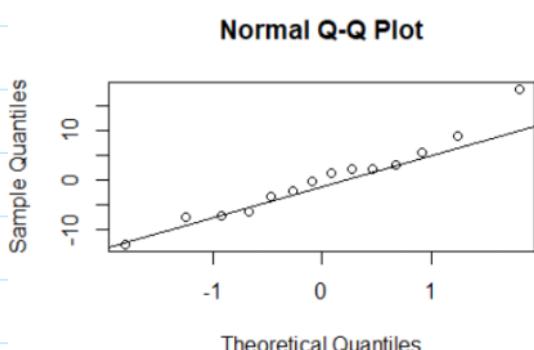
W = 0.96447, p-value = 0.7954

$> 0,05$

O erro é mt grande.

Não arriscamos rejeitar H_0 .

\therefore Os erros seguem distribuição Normal.



\Rightarrow Autocorrelação

lag Autocorrelation D-W Statistic p-value
1 -0.405677 2.757817 0.114 $> 0,05$
Alternative hypothesis: rho $\neq 0$

$H_0: \rho = 0$
 $H_1: \rho \neq 0$

Decisão: Não se Rejeita $H_0: \rho = 0 \rightarrow$ Não há correlação entre os erros

$$\rho = 0$$

Não São autocorr.

Os erros ϵ não são autocorrelacionados como se pretende.

Teste de Durbin Watson para avaliar a existência de autocorrelação dos resíduos, de ordem 1 i.e $t=pe$

$H_0: \rho = 0$ vs $H_1: \rho \neq 0$
durbinWatsonTest

$$\epsilon_i = \rho \epsilon_{i-1}$$

Só $\rho \neq 0$ são correlacionados

Conclusão: Os pressupostos estão validados
O modelo é válido para inferia.

Previsão

24 de maio de 2021 18:10

Prever com base no modelo obtido:

→ para um ind. que dorme em média 6 h:

$$\hat{y} = 6.233 + 1.091 \cdot \text{horas de sono}$$

$$\hat{y} = 6,233 + 1,091 \times 6 = 12,779 \text{ valores}$$

espera - se que tenha, em média,
12,779 valores, aprox. 13.

→ para um ind. que tem média de 13 valores:

$$13 = 6,233 + 1,091 \times X$$

$$X = \frac{13 - 6,233}{1,091} = 6,20$$

E' de esperar que durma cerca de
6,2 horas, em média, por noite.