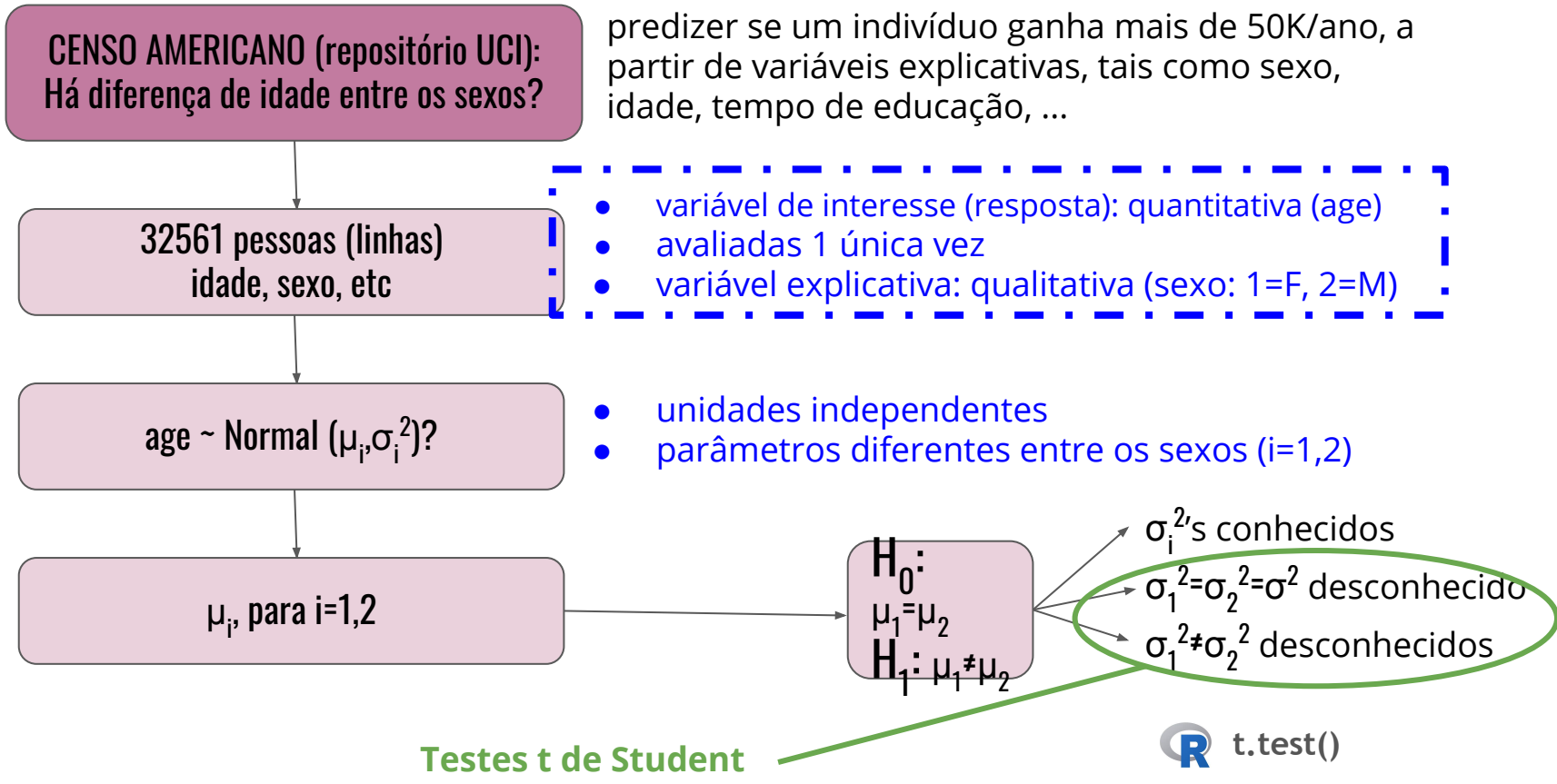
TESTES DE HIPÓTESES

5.5

DUAS POPULAÇÕES

5.5.3

Teste para a média de duas populações Normais (teste t de Student)



Teste t de Student: variâncias iguais

Pergunta: Há diferença entre as idades dos homens e das mulheres?

- X: idade do indivíduo
- X_{ij} ~ Normal (μ_i, σ^2), $E(X)=\mu_i$, $V(X)=\sigma^2$
  i = 1 (F), 2 (M)
  j = 1, ..., n=30162, independentes

(completos)

suposições:
- ❖ **variâncias iguais (homocedasticidade)**
- ❖ **distribuição normal**
- ❖ **independência**

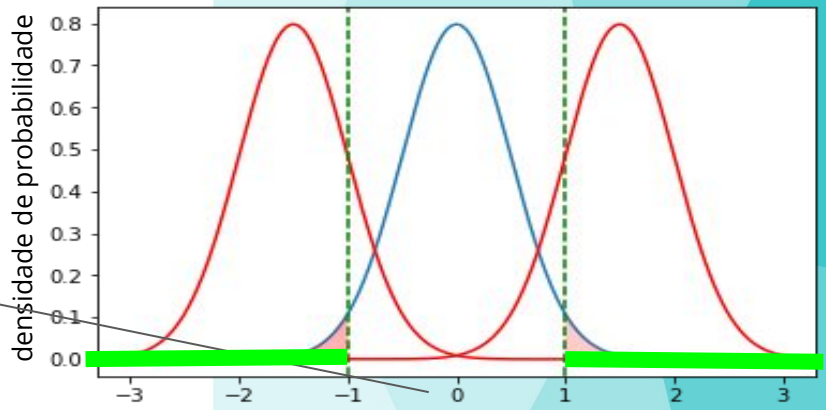
H_0: $\mu_1 = \mu_2$
H_a: $\mu_1 \neq \mu_2$

teste bicaudal

Estatística do teste

$$\frac{\bar{X}_1 - \bar{X}_2}{\sqrt{S_p^2(1/n_1 + 1/n_2)}},$$

$$S_p^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-1}$$



região de rejeição

$$\frac{\bar{X}_1 - \bar{X}_2}{\sqrt{S_p^2(1/n_1 + 1/n_2)}} \overset{\text{sob } H_0}{\sim} t_{n_1+n_2-2}$$

H_0: $\mu_1 = \mu_2$
H_a: $\mu_1 \neq \mu_2$

	age					
	count	min	max	median	mean	std
sex						
Female	9782	17	90	35	36.883459	13.532427
Male	20380	17	90	38	39.184004	12.873243

fixa-se α = 0,05
p-valor< 0,0001
Decido por H_a, pois p-valor < α

stats.ttest_ind(dataF,dataM, equal_var = True)

Decisão: em média, as idades de homens e mulheres não são iguais a um nível de significância de 5%

MAS 2 ANOS DE ≠ É IMPORTANTE? TAMANHO AMOSTRAL?

Verificação das suposições

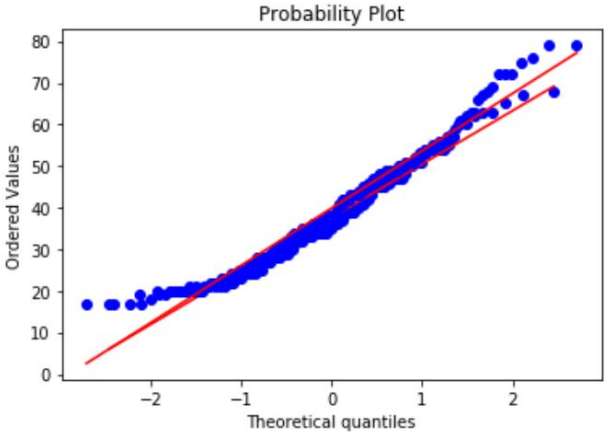
H_0: $\sigma_1^2 = \sigma_2^2$
H_a: $\sigma_1^2 \neq \sigma_2^2$

p-valor< 0,0001
Decido por H_a, pois p-valor < α

stats.levene()
stats.bartlett()

H_0: idade tem distribuição normal
H_a: idade não tem distribuição normal

stats.probplot()
stats.shapiro()
stats.kstest()



p-valor< 0,0001
Decido por H_a, pois p-valor < α

Teste t de Student: variâncias diferentes

Pergunta: Há diferença entre as idades dos homens e das mulheres?

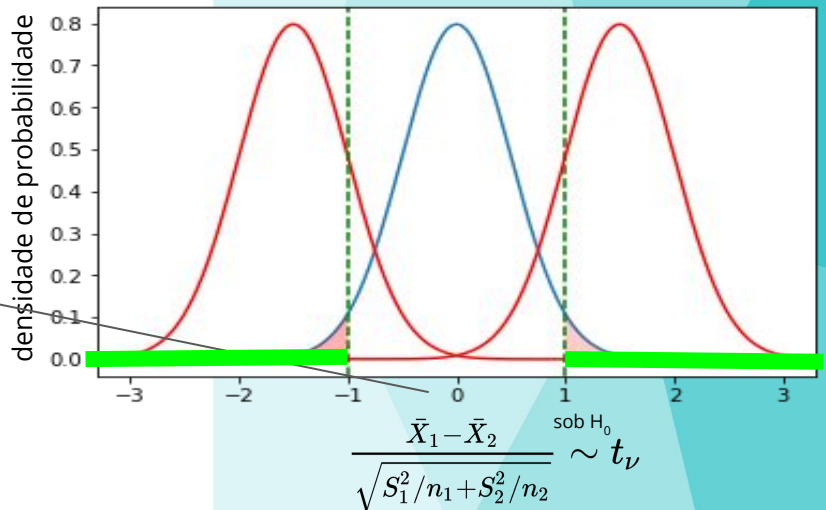
- X: idade do indivíduo
- X_{ij} ~ Normal (μ_i,σ_i^2), E(X)=μ_i, V(X)=σ_i^2
  i = 1 (F), 2 (M)
  j = 1, ..., n=32561, independentes

H_0: $\mu_1 = \mu_2$
H_a: $\mu_1 \neq \mu_2$

teste bicaudal

Estatística do teste

$$\frac{\bar{X}_1 - \bar{X}_2}{\sqrt{S_1^2/n_1 + S_2^2/n_2}}$$



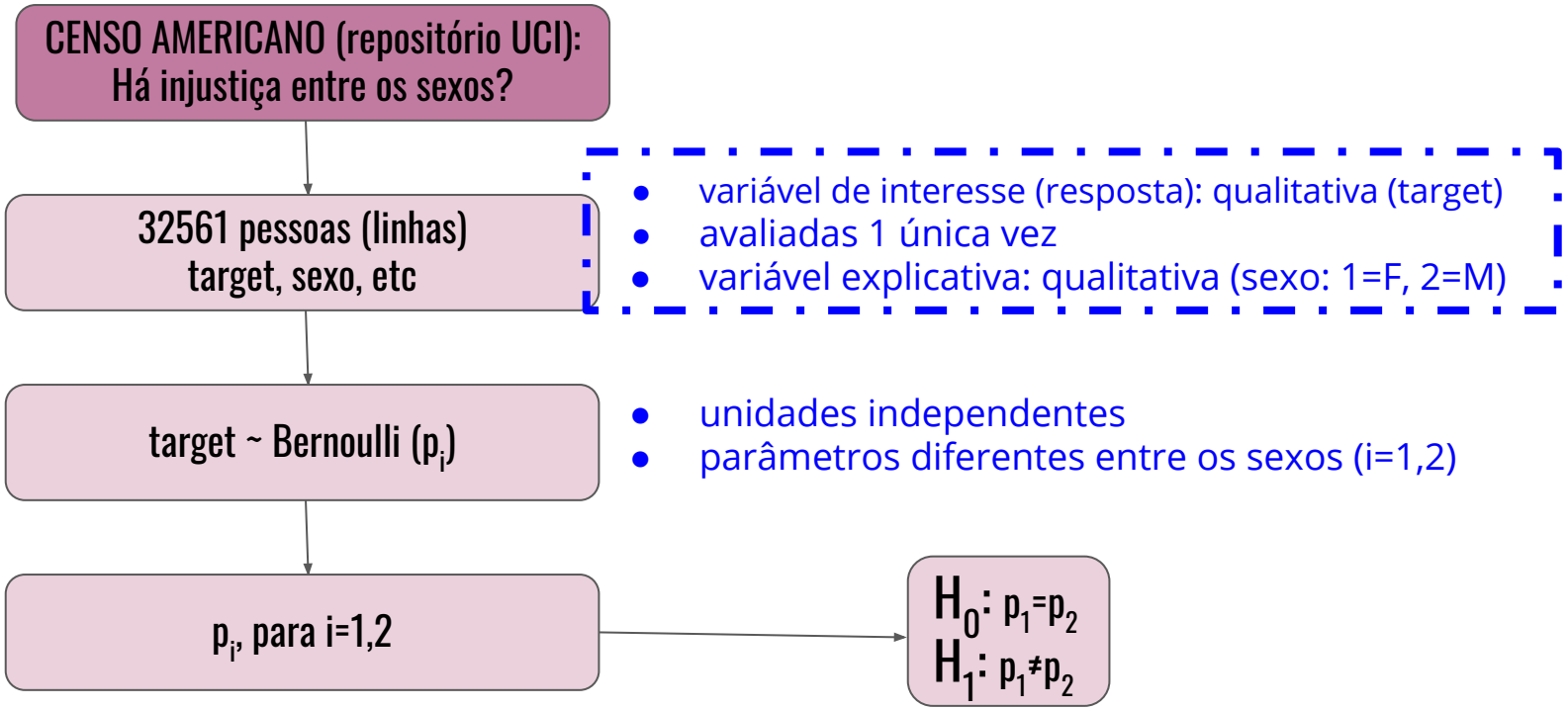
stats.ttest_ind(dataF,dataM, equal_var = **False**)

Testes de Hipótese frequentes
2. Sobre os parâmetros

$H_0 : \mu_1 = \mu_2$

H_1	*Caso*	*Estatística de teste*	*Rejeitar H_0 se:*
$\mu_1 - \mu_2 > 0$ $\mu_1 - \mu_2 < 0$ $\mu_1 - \mu_2 \neq 0$	1. σ_1 e σ_2 são conhecidas, as amostras são independentes e cada uma das populações tem distribuição normal ou os tamanhos de amostra n_i são suficientemente grandes	$Z = \dfrac{\overline{X} - \overline{Y}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$	$Z > z_{1-\alpha}$ $Z < z_\alpha = -z_{1-\alpha}$ $\lvert Z\rvert > z_{1-\alpha/2}$
$\mu_1 - \mu_2 > 0$ $\mu_1 - \mu_2 < 0$ $\mu_1 - \mu_2 \neq 0$	2. σ_1 e σ_2 são desconhecidas porém iguais, as amostras são independentes e as populações têm distribuição normal	$T = \dfrac{\overline{X} - \overline{Y}}{\sqrt{S_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$ Con $S_p^2 = \dfrac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}$ e a t é com n_1+n_2-2 g.l.	$T > t_{1-\alpha}$ $T < t_\alpha = -t_{1-\alpha}$ $\lvert T\rvert > t_{1-\alpha/2}$
$\mu_1 - \mu_2 > 0$ $\mu_1 - \mu_2 < 0$ $\mu_1 - \mu_2 \neq 0$	3. σ_1 e σ_2 são desconhecidas porém diferentes, as amostras são independentes e as populações têm distribuição normal	$T = \dfrac{\overline{X} - \overline{Y}}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$ e a t é com v g.l. con $v = \dfrac{\left(S_1^2/n_1 + S_2^2/n_2\right)^2}{\frac{(S_1^2/n_1)^2}{n_1+1} + \frac{(S_2^2/n_2)^2}{n_2+1}} - 2$	$T > t_{1-\alpha}$ $T < t_\alpha = -t_{1-\alpha}$ $\lvert T\rvert > t_{1-\alpha/2}$
$\mu_1 - \mu_2 > 0$ $\mu_1 - \mu_2 < 0$ $\mu_1 - \mu_2 \neq 0$	4. σ_1 e σ_2 são desconhecidas, as amostras são independentes e têm tamanhos suficientemente grandes	$Z = \dfrac{\overline{X} - \overline{Y}}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$	$Z > z_{1-\alpha}$ $Z < z_\alpha = -z_{1-\alpha}$ $\lvert Z\rvert > z_{1-\alpha/2}$

Teste para a comparação de 2 proporções



Teste exato: teste qui-quadrado para independência ou homogeneidade de distribuições (Bernoulli)

$H_0: p_1 = p_2$
$H_1: p_1 \neq p_2$

target	<=50K	>50K	All
sex			
Female	8670 (89%)	1112 (11%)	9782
Male	13984 (69%)	6396 (31%)	20380
All	22654	7508	30162

$\hat{p}_1$

$\hat{p}_2$

p-valor < 0,0001:

Decisão: distribuição de 'target' não é a mesma entre homens e mulheres a um nível de 5%

chisq.test

scipy.stats.chi2_contingency

Teste aproximado (pelo TCL)

H_0: $p_1 = p_2$
H_1: $p_1 \neq p_2$

target	<=50K	>50K	All
sex			
Female	8670 (89%)	1112 (11%)	9782
Male	13984 (69%)	6396 (31%)	20380
All	22654	7508	30162

$\hat{p}_1$

$\hat{p}_2$

$$\hat{p}_i \approx N(p_i, p_i(1-p_i)/n_i)$$

statsmodels.stats.weightstats.ztest

Testes de Hipótese frequentes

2. Sobre os parâmetros

$\boldsymbol{H_0 : \sigma_1^2 = \sigma_2^2}$

H_1	*Caso*	*Estatística de teste*	*Rejeitar H_0 se:*
$\sigma_1^2 > \sigma_2^2$ $\sigma_1^2 < \sigma_2^2$ $\sigma_1^2 \neq \sigma_2^2$	As amostras são independentes e as populações têm distribuição normal	$F = \frac{S_1^2}{S_2^2}$ e a F é com $n_1 - 1$ e $n_2 - 1$ g.l.	$F > F_{1-\alpha}$ $F < F_{\alpha}$ $F < F_{\alpha/2}$ o $F > F_{1-\alpha/2}$

$\boldsymbol{H_0 : p_1 = p_2}$

H_1	*Caso*	*Estatística de teste*	*Rejeitar H_0 se:*
$p_1 > p_2$ $p_1 < p_2$ $p_1 \neq p_2$	As amostras são independentes e os tamanhos de amostra n_i são suficientemente grandes	$Z = \frac{\bar{p}_1 - \bar{p}_2}{\sqrt{\bar{p}\bar{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$ con $\bar{p} = (n_1\bar{p}_1 + n_2\bar{p}_2)/(n_1 + n_2)$	$Z > z_{1-\alpha}$ $Z < z_{\alpha} = -z_{1-\alpha}$ $\lvert Z\rvert > z_{1-\alpha/2}$

TEOREMA CENTRAL DO LIMITE

Effects of Departures from Normality

If the probability distributions of Y are not exactly normal but do not depart seriously, the sampling distributions of b_0 and b_1 will be approximately normal, and the use of the t distribution will provide approximately the specified confidence coefficient or level of significance. Even if the distributions of Y are far from normal, the estimators b_0 and b_1 generally have the property of *asymptotic normality*—their distributions approach normality under very general conditions as the sample size increases. Thus, with sufficiently large samples, the confidence intervals and decision rules given earlier still apply even if the probability distributions of Y depart far from normality. For large samples, the t value is, of course, replaced by the z value for the standard normal distribution.

Fonte: Kutner et al. Applied Linear Statistical Models