

Estatística para Ciências de Dados

Aula 5: Testes de hipóteses e inferência baseada em simulação

Mariana Cúri
ICMC/USP
mcuri@icmc.usp.br



Conteúdo

1. Motivação

- a. Artigos Sildenafil
- b. Definição do modelo probabilístico e hipóteses do teste

2. Testes de hipóteses

- a. Conceitos básicos: um exemplo para proporção
- b. Cálculo do tamanho amostral
- c. Teste t de Student

3. Inferência baseada em simulação

Motivação: estudo 1 Sildenafil

BMC Urology

Research article

Sildenafil (Viagra) for male erectile dysfunction: clinical trial reports

RA Moore*², JE Edwards¹ and HJ McQuay¹

Address: ¹Pain Research and Nuffield Department of Anaesthetics, University of Oxford, Oxford Oxford OX3 7LJ, UK and ²Pain Research and Nuffield Department of Anaesthetics, University of Cambridge, Cambridge CB2 3RQ, UK

Table 1: At least 60% of attempts at sexual intercourse successful

Dosing (mg)	Number of trials	Number (%) with outcome			
		Sildenafil	Placebo	Relative benefit (95% CI)	NNT (95% CI)
25	3	88/312 (28)	43/426 (10)	3.0 (2.1 to 4.2)	5.5 (4.2 to 8.1)
50	5	216/511 (42)	62/607 (10)	4.3 (3.3 to 5.6)	3.1 (2.7 to 3.7)
100	5	223/506 (44)	62/607 (10)	4.4 (3.4 to 5.8)	3.0 (2.6 to 3.5)
200	2	93/191 (49)	19/181 (10)	4.5 (2.9 to 7.1)	2.6 (2.2 to 3.4)
Dose optimised	3	183/379 (48)	43/376 (11)	4.2 (3.1 to 5.6)	2.7 (2.3 to 3.2)

Abstract

Background: Evaluation of company clinical trial reports could provide information for meta-analysis at the commercial introduction of a new technology.

Methods: Clinical trial reports of sildenafil for erectile dysfunction from September 1997 were used for meta-analysis of randomised trials (at least four weeks duration) and using fixed or dose optimisation regimens. The main outcome sought was an erection, sufficiently rigid for penetration, followed by successful intercourse, and conducted at home.

Results: Ten randomised controlled trials fulfilled the inclusion criteria (2123 men given sildenafil and 1131 placebo). NNT or NNH were calculated for important efficacy, adverse event and discontinuation outcomes. Dose optimisation led to at least 60% of attempts at sexual intercourse being successful in 49% of men, compared with 11% with placebo; the NNT was 2.7 (95% confidence interval 2.3 to 3.3). For global improvement in erections the NNT was 1.7 (1.6 to 1.9). Treatment-related adverse events occurred in 30% of men on dose optimised sildenafil compared

with placebo (20%). Sildenafil dose optimisation gave efficacy equivalent and adverse events equivalent to the lowest fixed doses.

of clinical trial reports available at the time of licensing agreed with later randomised trials and patients. Making reports submitted for marketing approval provide better information when it was most needed, and would improve evaluation of new technologies.

Motivação: estudo 2 Sildenafil

Clinical Urology

International Braz J Urol

Vol. 31 (4): 342-355, July - August, 2005

Official Journal of the Brazilian Society of Urology

EFFICACY, SAFETY AND TOLERABILITY OF SILDENAFIL IN BRAZILIAN HYPERTENSIVE PATIENTS ON MULTIPLE ANTIHYPERTENSIVE DRUGS

DENILSON C. ALBUQUERQUE, LINEU J. MIZIARA, JOSE F. K. SARAIVA, ULISSES S. RODRIGUES, ARTUR B. RIBEIRO, MAURICIO WAJNGARTEN

Department of Cardiology (DCA), State University of Rio de Janeiro, Department of Cardiology (LJM), Federal University of Uberlandia, Minas Gerais, Department of Cardiology (JFKS), Pontifical Catholic University, Campinas, Sao Paulo, Department of General Practice (USR), Salgado Filho Hospital, Rio de Janeiro, Department of Nephrology (ABR), Federal University of Sao Paulo, UNIFESP, CardioGeriatry Service (MW), Institute of Heart, INCOR, Sao Paulo, Brazil

The proportion of successful attempts at intercourse was compared between the two groups using a generalized estimation equation model assuming a uniform structure for the correlation. All hypothesis testing considered a p value ≤ 0.05 as statistically significant.

The analysis of event logs demonstrated statistically significant differences between the two groups in the proportions of successful attempts at sexual intercourse. Among patients treated with sildenafil, successful attempts were reported in 54%, 61% and 73% of the times after 2, 4 and 8 weeks of treatment. Among patients that took the placebo, these same proportions were 13%, 20% and 29% ($p < 0.0001$ for the comparison between groups at each time point).

Motivação: definição do modelo e das hipóteses

Grupo	Paciente	Semana	Tentativa	Resultado
Sildenafil	1	1	1	1
	1	1	2	1
	1	2	1	0
	1	2	2	1
	1	2	3	0
	1	3	1	1
	1	3	2	1
	1	4	1	1
	2	1	1	0
	2	1	2	0
	2	2	1	1
	2	3	1	0
	2	4	1	1
	3	1	1	1
	3	1	2	0

Bernoulli(p_g), dependência dentro de indivíduo

Hipóteses: $p_{sil}=p_{pl}$ vs $p_{sil}>p_{pl}$? $g=sil, pl$

Poisson(λ_g)?

% \sim Normal(μ_g, σ^2)?

\sim Beta(α, β)?

categorizar: $\geq 60\%$ ou $<60\%$?

\sim Bernoulli(θ_g), independência

possíveis modelos probabilísticos

hipóteses do teste

Testes de hipóteses

Exemplo:

Teste para a proporção de uma população

Hipóteses do teste:

H_0 : Sildenafil é eficaz (hipótese nula)

H_a ou H_1 : Sildenafil não é eficaz (hipótese alternativa)

hipótese simples:
um único valor para o parâmetro

$$H_0 \cap H_1 = \emptyset \text{ e } H_0 \cup H_1 = \Theta$$

$$H_0: p = 0,5$$

$$H_a: p < 0,5$$

hipótese
composta:
mais de um
valor para o
parâmetro

- X: pelo menos 60% das tentativas bem sucedidas
(para tornar didático, suponha apenas o grupo Sildenafil com dose otimizada)
- $X \sim \text{Bernoulli}(p)$, $i = 1, \dots, n=379$, independentes, $E(X)=p$, $V(X)=p(1-p)$
p: proporção de pessoas (na população) com pelo menos 60% das tentativas bem sucedidas

$$X_1 = 1,$$

$$X_2 = 0,$$

$$X_3 = 1, \dots$$

$$\hat{p}_{obs} = \frac{183}{379} = 0,483$$

Como decido?

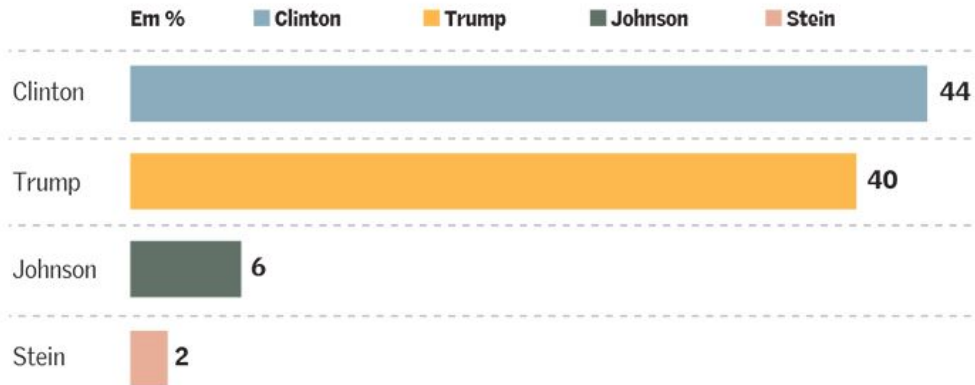
no teste, uso as observações da
amostra para decidir entre H_0 e H_a

Testes de hipóteses: (p x \hat{p})

Um parênteses em nosso exemplo: p vs \hat{p}

Reta de chegada

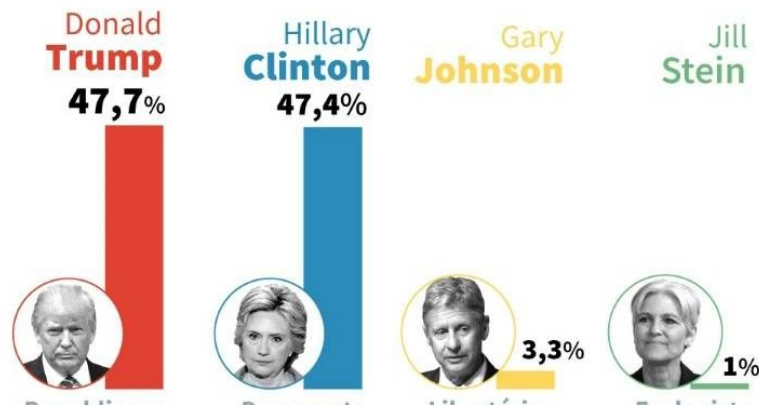
Última pesquisa do WSJ e a rede NBC News antes da eleição



Fonte: Pesquisa WSJ/NBC News com eleitores realizada entre 3 e 5/11

THE WALL STREET JOURNAL

Porcentagem dos votos Resultados às 10:30h (Brasília)



Testes de hipóteses: proporção de uma população

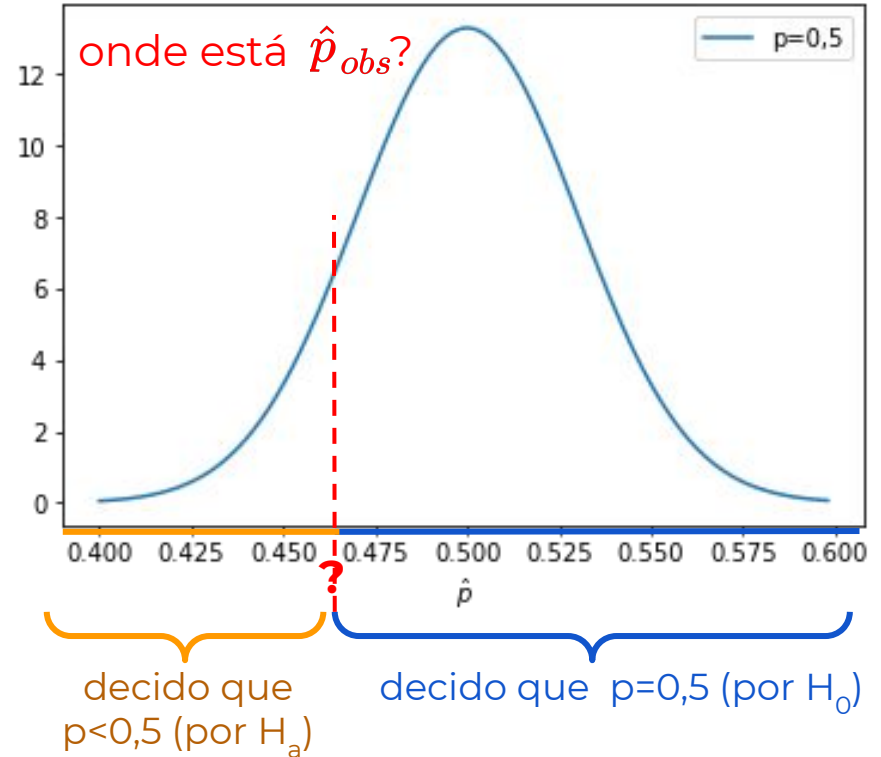
$$H_0: p = 0,5$$

$$H_a: p < 0,5 \quad \leftarrow \text{teste unicaudal}$$

Estatística do teste e distribuição amostral (TCL):

$$\frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \approx N(0, 1), \text{ pois } \hat{p} \approx N\left(p, \frac{p(1-p)}{n}\right)$$

		decisão	
		H_0	H_a
verdade	H_0	sem erro	Erro Tipo I
	H_a	Erro Tipo II	sem erro

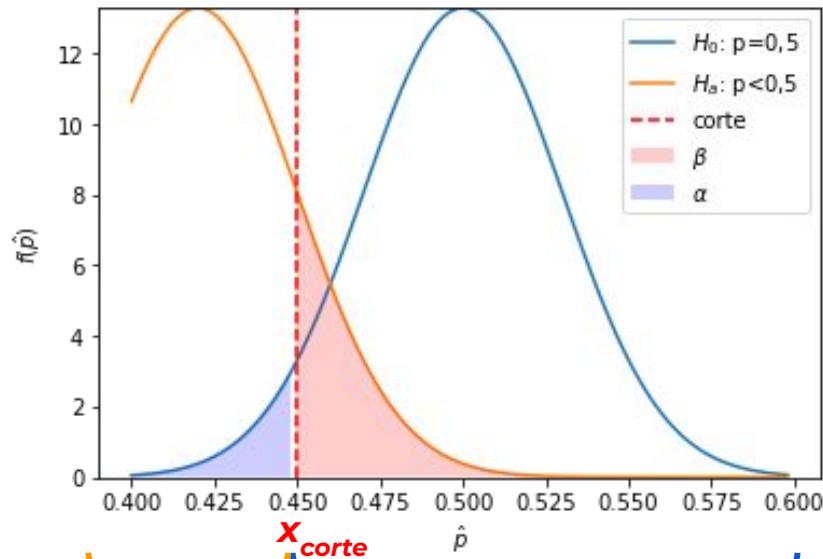


Testes de hipóteses: proporção de uma população

$P(\text{erro tipo I}) = P(\text{decidir por } H_a \text{ sendo } H_0 \text{ verdadeira}) = \alpha$

$P(\text{erro tipo II}) = P(\text{decidir por } H_0 \text{ sendo } H_a \text{ verdadeira}) = \beta$

↓ dada a amostra,
↓ não dá para diminuir
ambos simultaneamente



Fixo α em um valor pequeno: 0,01? 0,05? 0,1?

$P(\text{erro tipo I}) = \alpha = 0,01$: **nível de significância**

Se H_0 é verdadeira, então: $\hat{p} \approx N\left(0,5; \frac{0,5 \cdot (1-0,5)}{379}\right)$

$$P(\hat{p} \leq x_{corte} \mid p = 0,5) = 0,01$$

$$x_{corte} = 0,44$$

Região crítica

$$R_c = \{\hat{p} \leq 0,44\}$$

Região de aceitação

$$R_a = \{\hat{p} > 0,44\}$$

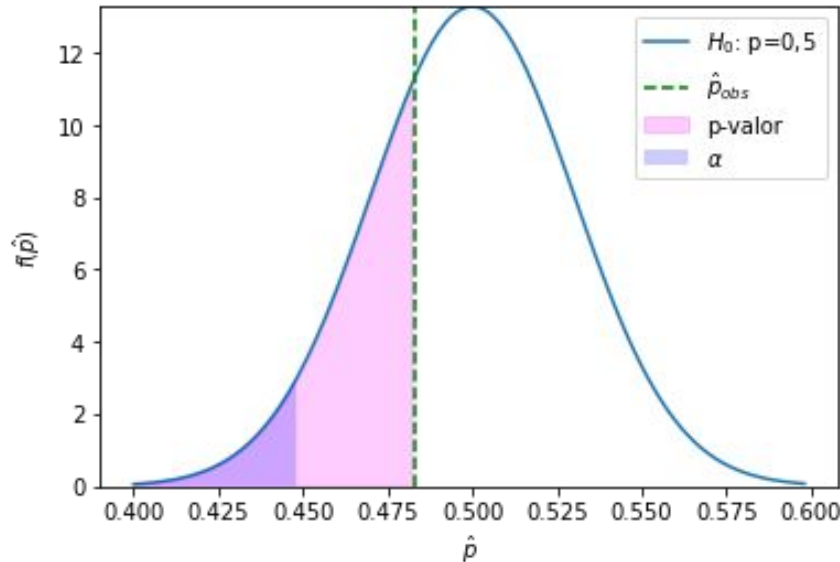
decido que $p < 0,5$ (por H_a)

decido que $p = 0,5$ (por H_0)

Como $\hat{p}_{obs} = 0,483 \in R_a$, decido por H_0
para $\alpha = 1\%$

Testes de hipóteses: proporção de uma população

Outra forma de tomar a decisão



Nível descritivo ou p-valor

probabilidade de se obter uma estatística de teste igual ou mais extrema que aquela observada na amostra, sob H_0

$$\begin{aligned} \text{p-valor} &= P(\hat{p} \leq 0,483 \mid p = 0,5) \\ &= 0,25 > \alpha : \text{decide por } H_0 \end{aligned}$$

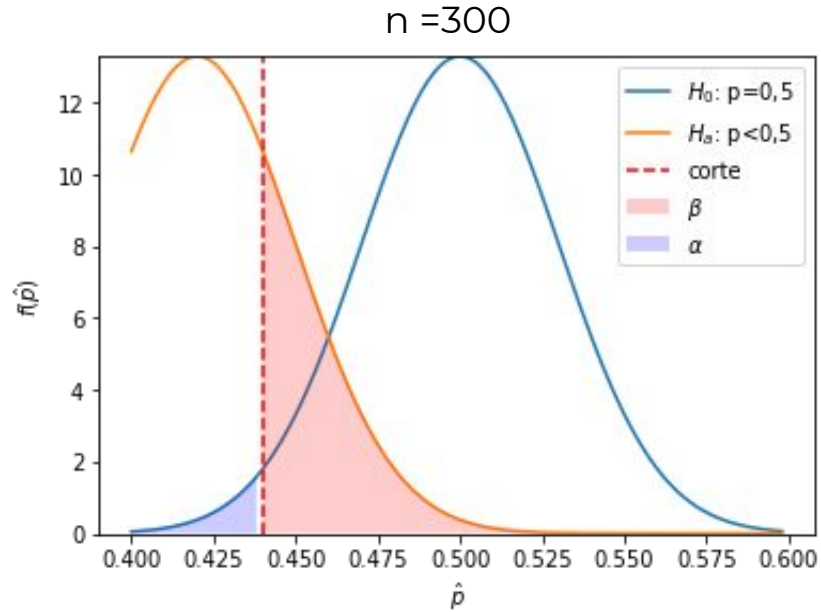
$$\text{p-valor} \begin{cases} > \alpha, \text{ decide por } H_0 \\ < \alpha, \text{ decide por } H_a \end{cases}$$

Testes de hipóteses

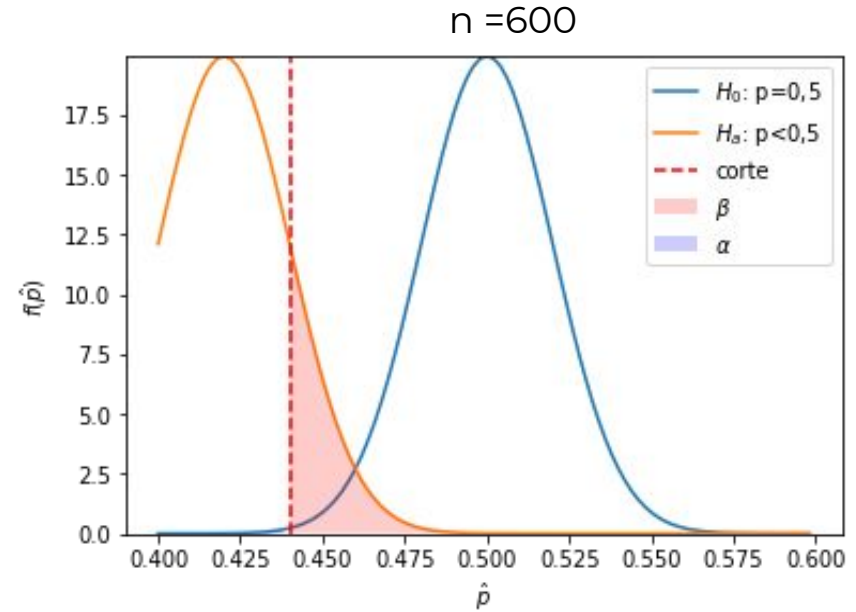
Passos de um teste de hipóteses

- Especificar (em termos dos parâmetros) as hipóteses H_0 e H_a
- Especificar a estatística do teste e sua distribuição, sob H_0
- Fixar o nível de significância do teste (α)
- Calcular o p-valor (ou a região crítica do teste)
- Decidir entre H_0 e H_a , comparando o p-valor com α (ou verificando se a estatística do teste pertence ou não à região crítica)

Testes de hipóteses: tamanho amostral

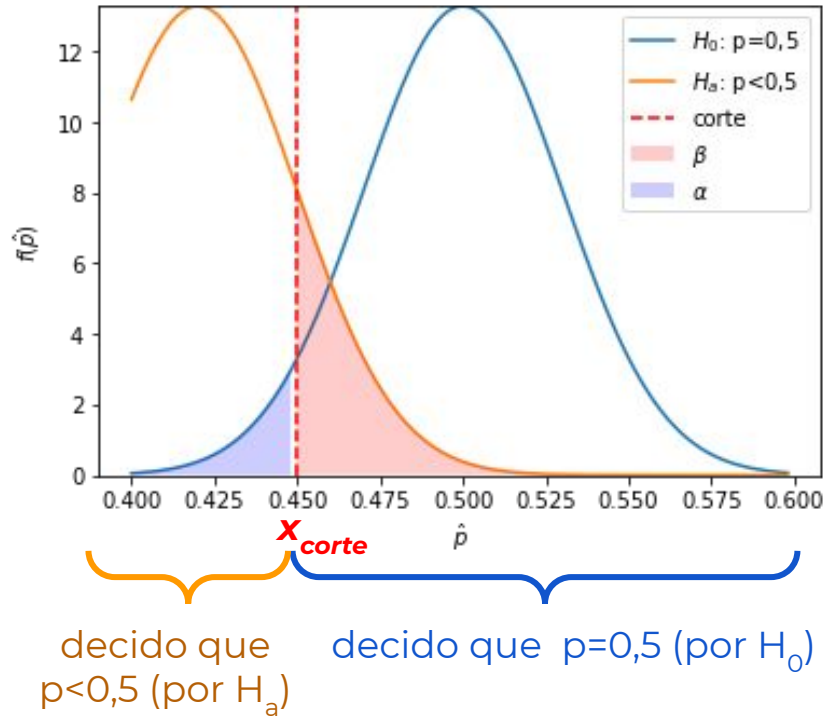


$$\hat{p} \approx N\left(p, \frac{p(1-p)}{n}\right)$$



equilibrar α e β variando n

Testes de hipóteses: tamanho amostral



Qual o tamanho amostral para
testar a eficácia do Sildenafil?

dados necessários:

$$H_0: p = p_0$$

$$H_a: p < p_0$$

unicaudal

teste

nível de significância, α
poder do teste, $\pi = 1 - \beta$

$$p_0$$

$\varepsilon = p - p_0$ *diferença a ser detectada*
 $\text{Var}(X) = p(1-p)$ *amostra piloto*

Testes de hipóteses: tamanho amostral

$$\alpha = P(\hat{p} \leq x_{corte} \mid p = p_0) \\ = P\left(Z \leq \frac{x_{corte} - p_0}{\sqrt{p_0(1-p_0)/n}}\right)$$

$$x_{corte} = p_0 + z_\alpha \sqrt{p_0(1-p_0)/n}$$

$$\pi = 1 - \beta = 1 - P(\hat{p} > x_{corte} \mid p) \\ = P(\hat{p} \leq x_{corte} \mid p)$$

$$x_{corte} = p + z_\pi \sqrt{p(1-p)/n}$$

$$n = \left(\frac{z_\alpha \sqrt{p_0(1-p_0)} - z_\pi \sqrt{p(1-p)}}{\epsilon} \right)^2$$

dados necessários:

$$H_0: p = p_0$$

$$H_a: p < p_0$$

unicaudal

teste

nível de significância, $\alpha = 0,05$, $z_\alpha = -1,64$

poder do teste, $\pi = 1 - \beta = 0,80$, $z_\pi = 1,28$

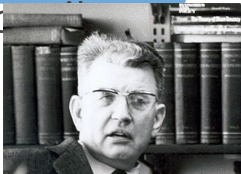
$$p_0 = 0,5$$

$$\epsilon = p - p_0$$

diferença a ser detectada

$$\text{Var}(X) = p(1-p)$$

"If you torture the data long enough, it will confess."



Ronald Coase
1910-2013

$\epsilon = 0,01 \rightarrow$ da ordem de 15mil

$\epsilon = 0,05 \rightarrow$ da ordem de 600

Teste de hipóteses: teste t de Student

Pergunta: Há diferença entre alunos de escolas privadas e públicas no desempenho no ENEM?

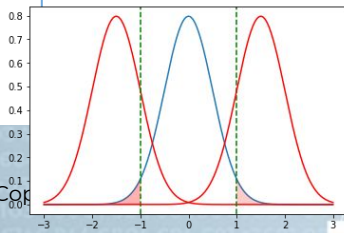
Tipo de escola	Média no ENEM	Desvio Padrão	nº de escolas
Estadual	497.1	30.7	1085
Privada	545.8	29.3	1816
Total Geral	527.6	38.1	2901

- X : média dos alunos da escola no ENEM
- $X_{ij} \sim \text{Normal}(\mu_j, \sigma^2)$, $E(X) = \mu_j$, $V(X) = \sigma^2$
 $i = 1, \dots, n = 2901$, independentes
 $j = 1$ (pública), 2 (privada)

$$H_0: \mu_1 = \mu_2$$

$$H_a: \mu_1 \neq \mu_2$$

teste bicaudal



Estatística do teste

$$\frac{\bar{X}_1 - \bar{X}_2}{\sqrt{S_p^2(1/n_1 + 1/n_2)}},$$
$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 1}$$

fixa-se $\alpha = 0,05$ p-valor $< 0,0001$
Decido por H_a , pois p-valor $< \alpha$,
em média, as escolas não têm o
mesmo resultado médio no ENEM


Inferência baseada em simulação: bootstrap

- Método para simular distribuição amostral, selecionando aleatoriamente múltiplas amostras (com reposição) da amostra original
- Publicado por Bradley Efron, em 1979, inspirado em um outro método (Jackknife)
- Estimar erro padrão, vício, construir IC, etc
- Necessita, basicamente, de uma boa amostra, apenas
- Há dois tipos: paramétrico e não paramétrico
- Paramétrico: sabe-se a distribuição da amostra e esta é usada apenas para estimar os parâmetros dessa distribuição; as subamostras são geradas a partir dessa classe de distribuição com as estimativas obtidas
- Não-paramétrico: a distribuição empírica é usada para a obtenção das subamostras

Inferência baseada em simulação: bootstrap

Algoritmo (não-paramétrico)

- $\mathbf{x} = (x_1, x_2, \dots, x_n)$ amostra original
- geram-se B sub-amostras \mathbf{x}_i^* , $i = 1, 2, \dots, B$, também de tamanho n ,
selecionando-se n elementos de \mathbf{x} , com reposição
- para cada uma das B sub-amostra bootstrap geradas, calcula-se o estimador de interesse $\hat{\theta}_i^*, i = 1, 2, \dots, B$
- Com as B 's estimativas desse estimador, calcula-se a quantidade de interesse: erro padrão (desvio padrão), o vício, IC, etc

Para o [bootstrap paramétrico](#), substitui-se  por n observações geradas da classe de distribuições assumidas para amostra, usando-a para estimação dos parâmetros dessa distribuição