TESTES DE HIPÓTESES

CONCEITOS BÁSICOS

5.4

BMC Urology

BioMed Central

Research article

Sildenafil (Viagra) for male erectile dysfunction: a meta-analysis of clinical trial reports

RA Moore*[2], JE Edwards[1] and HJ McQuay[1]

Estudo 1 Sildenafil

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 11% on placebo; the NNH was 5.4 (4.3 to 7.3). All cause discontinuations were less frequent with sildenafil (10%) than with placebo (20%). Sildenafil dose optimisation gave efficacy equivalent to the highest fixed doses, and adverse events equivalent to the lowest fixed doses.

Conclusion: This review of clinical trial reports available at the time of licensing agreed with later reviews that had many more trials and patients. Making reports submitted for marketing approval available publicly would provide better information when it was most needed, and would improve evidence-based introduction of new technologies.

Estudo 2 Sildenafil

Clinical Urology

International Braz J Urol
Official Journal of the Brazilian Society of Urology

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

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

Dep... ...LJM), Fede... ...atholic Unive... ...Rio de Jane... ...eriatry

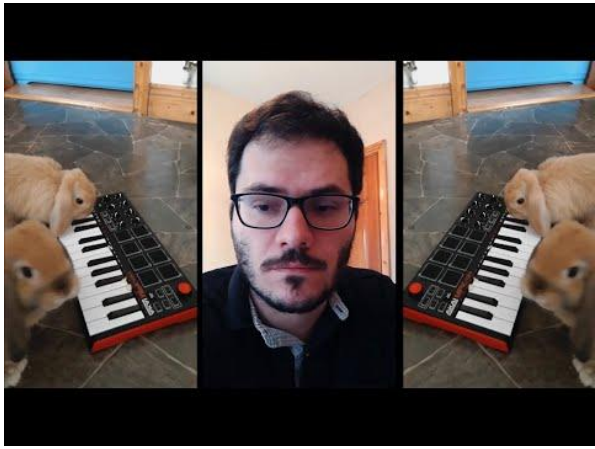
The proportion of successful attempts at intercourse was compared between the two groups using a generalized estimation equation model assuming a uniform structure for the correlation. All hypothesis testing considered a p value ≤ 0.05 as statistically significant.

Estudo 2 Sildenafil

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).

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
				0
				1
				1
				1



Summary Song #2 - Discrete Random Variables
by Rafael de Andrade Moral (https://www.youtube.com/watch?v=ZINXFoQMZVs)

Letra:
The binomial distribution is a sum of Bernoullis
But only if they're independent and they all have the same p

Y is a discrete proportion of N
Here the letter p means probability
It's easy to say the mean is np
And so the variance is np times one minus p

We got the Bernoulli
We got the Poisson
They're discrete
That's why I wrote this song
There's the binomial
But when Y is a count
e minus lambda lambda x over x factorial

You think it's simple but it's not there's a whole heap of assumptions
Events must be independent and same rate of occurrence

That's for the Poisson
That's right, it's a name
The variance and mean are exactly the same
We know that may be a bit way too strict
But this model is a pretty good start generally

The pmf has to sum to one
Or it will fail validation
Computing the Var
It's not that complex
It's just E of X squared minus the square of E of X

And I understand it may be confusing sometimes
Just remember that your model choice must reflect the true nature of Y

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)?

% ~ Normal(μ_g,σ^2)?

~ Beta(α,β)?

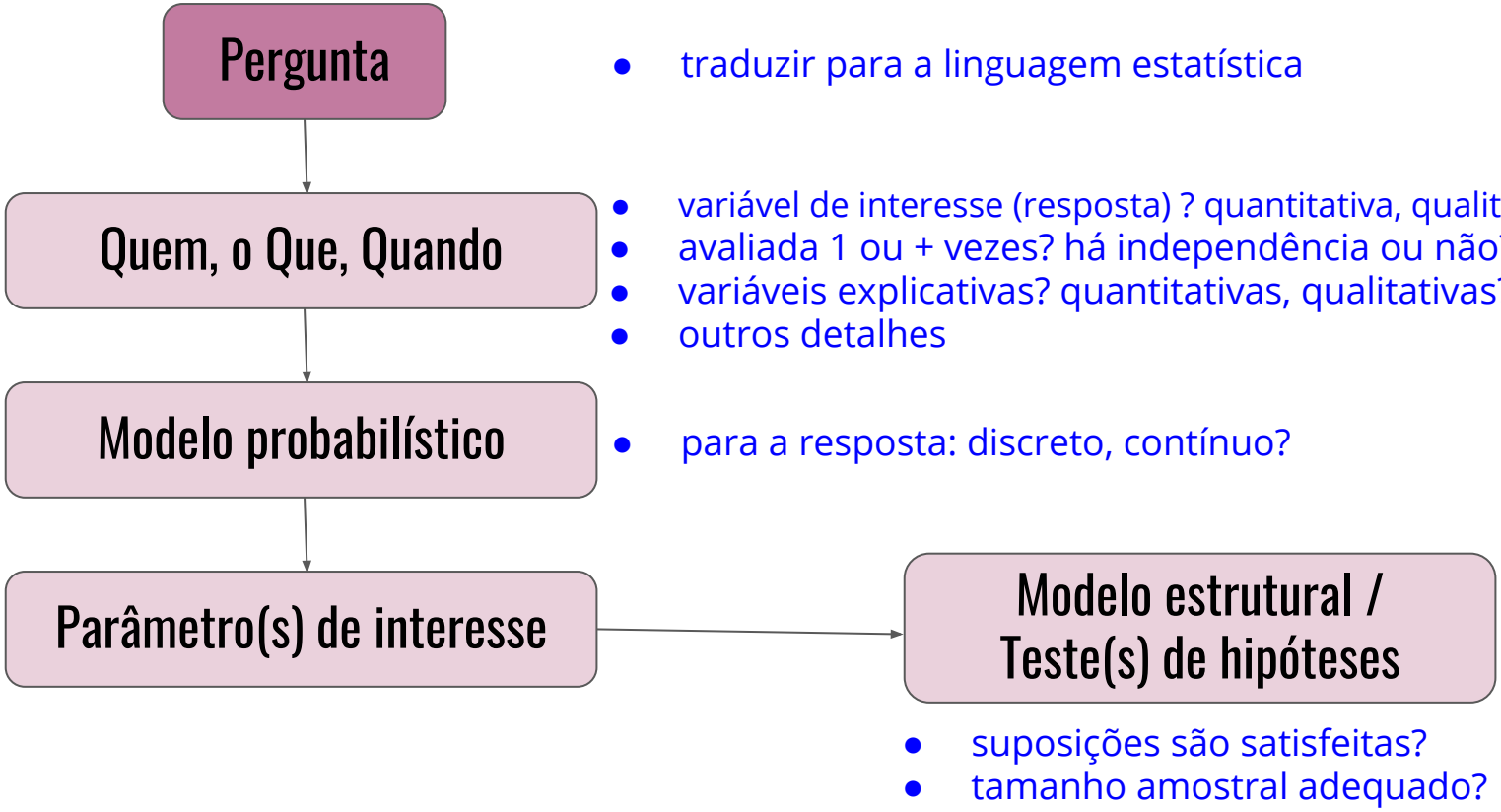
categorizar: ≥60% ou <60%?

~ Bernoulli(θ_g), independência

possíveis modelos probabilísticos

hipóteses do teste

Sequência de raciocínio para a inferência estatística



Estudo 1 Sildenafil: análise descritiva

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

Dosing (mg)	Number of trials	Number (%) with outcome		Relative benefit (95% CI)	NNT (95% CI)
		Sildenafil	Placebo		
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)

Estido 1 Sildenafil: testes de hipóteses

Pergunta de interesse:

Sildenafil é eficaz (hipótese nula)

Sildenafil não é eficaz (hipótese alternativa)

2 hipóteses complementares

Y_{gi}: indicadora de tentativa bem sucedida

$Y_{ig} \sim Bernoulli(p_{g(i?)})$

i: i-ésimo indivíduo do grupo g

g: 1=sil, 2=pl

dependentes no mesmo indivíduo

F_{gi}: porcentagem de tentativas bem sucedidas

$F_{ig} = \frac{Y_{gi}}{n_{gi}}$

n_{gi}: número de tentativas do indivíduo i do grupo g

independentes

X_{gi}: indicadora de $F_{gi} \geq 0.6$

$X_{gi} \sim Bernoulli(p_g)$

independentes

Tradução para a linguagem estatística

H_0: p = 0,5

H_a: p < 0,5

por questões didáticas: apenas grupo Sildenafil

X_i: indicadora de $F_i \geq 0.6$

$X_i \sim Bernoulli(p)$

independentes

Estudo 1 Sildenafil: análise descritiva

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

Dosing (mg)	Number of trials	Number (%) with outcome		Relative benefit (95% CI)	NNT (95% CI)
		Sildenafil	Placebo		
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)

Exemplo:

Teste para a proporção de uma população

$H_0 \cap H_1 = \varnothing$ e $H_0 \cup H_1 = \Theta$

H_0: p = 0,5

H_a: p < 0,5

hipótese simples:
um único valor para o parâmetro

hipótese composta:
mais de um valor para o parâmetro

- X: pelo menos 60% das tentativas bem sucedidas

- X ~ Bernoulli (p), i = 1, ..., 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

$X_1 = 1,$
$X_2 = 0,$
$X_3 = 1, \ldots$

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

Como decido?

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

Exemplo:

Teste para a proporção de uma população

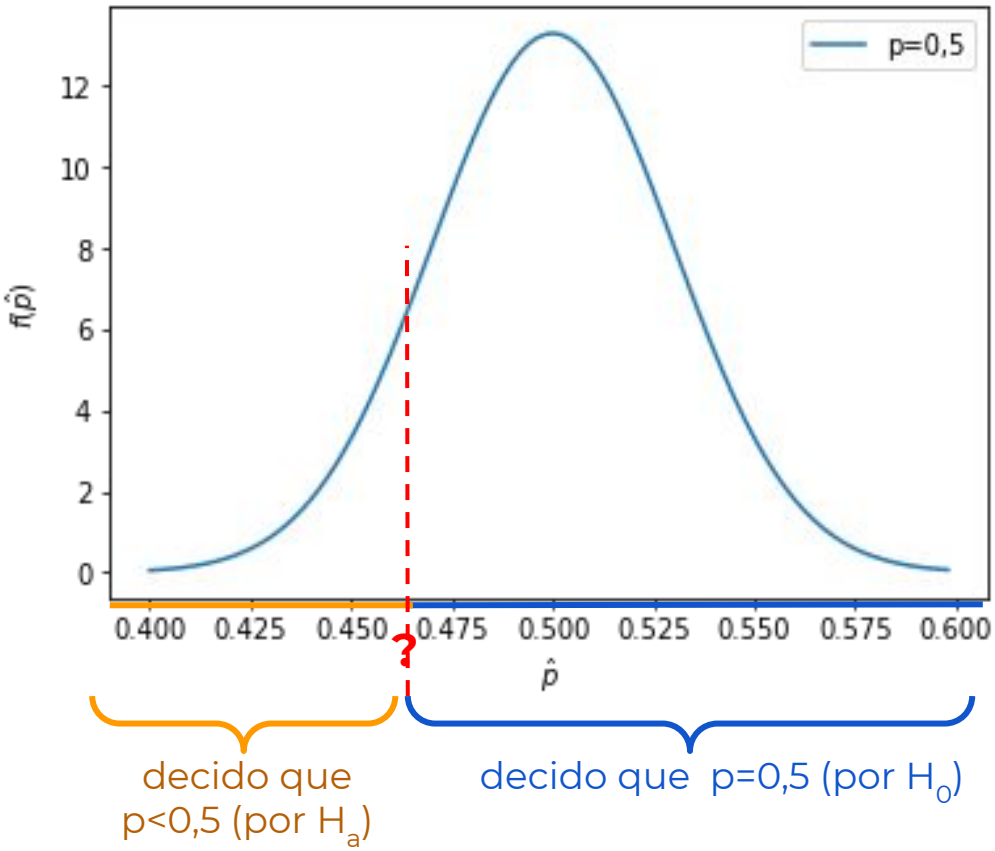
H_0: p = 0,5

H_a: p < 0,5 ← teste unicaudal (à esquerda)

Estatística do teste e distribuição amostral (TCL):

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

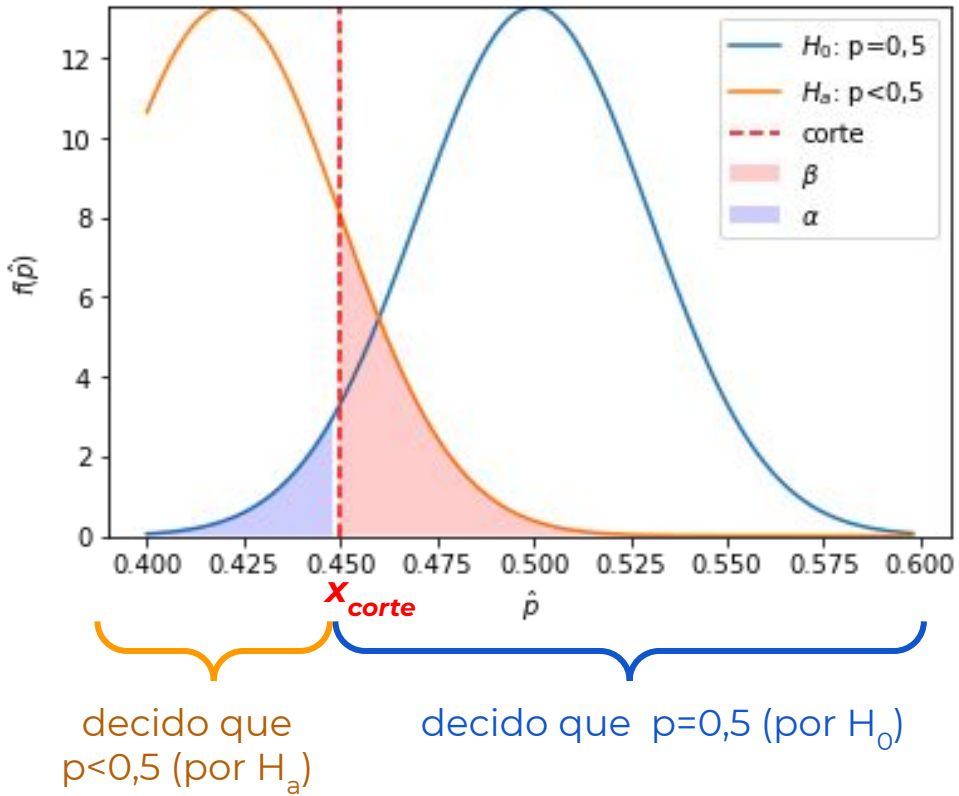
verdade	decisão: H_0	decisão: H_a
H_0	sem erro	Erro Tipo I
H_a	Erro Tipo II	sem erro

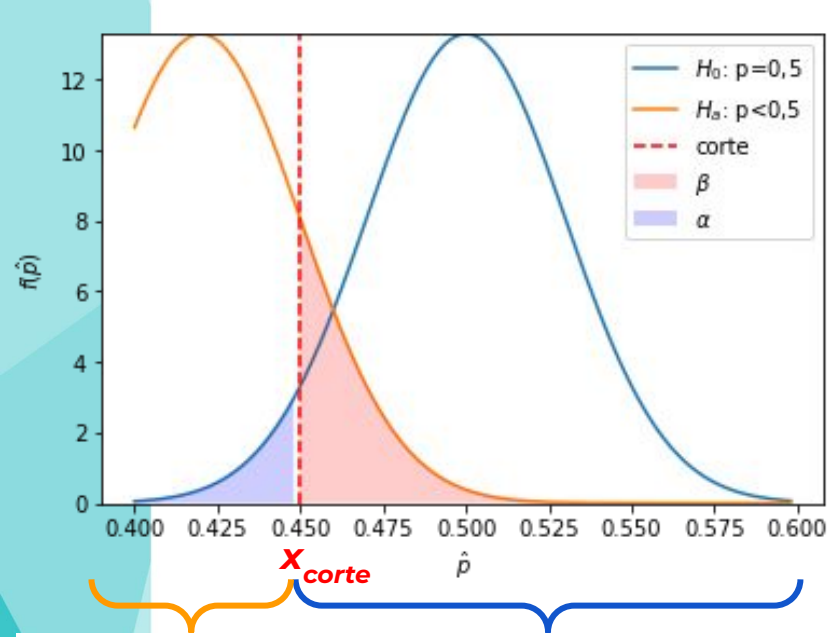

p=0,5
12
10
8
6
4
2
0
f(p̂)
0.400
0.425
0.450
0.475
0.500
0.525
0.550
0.575
0.600
p̂
?
decido que
p<0,5 (por H_a)
decido que p=0,5 (por H_0)

P(erro tipo I) = P(decidir por H_a sendo H_0 verdadeira) = **α**

P(erro tipo II) = P(decidir por H_0 sendo H_a verdadeira) = **β**

dada a amostra,
não dá para diminuir
ambos simultaneamente





decido que p<0,5 (por H_a)

decido que p=0,5 (por H_0)

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

P(erro tipo I) = **α** = 0,01**: nível de significância**

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

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

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

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

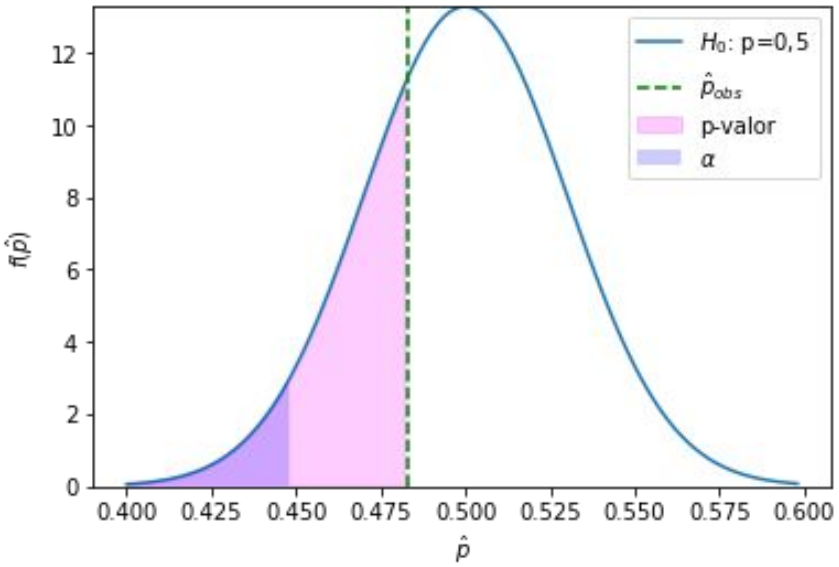
Região crítica

$$R_c = \{\hat{p} \leq 0,44\} \text{ ou } \frac{\hat{p} - 0,5}{\sqrt{0,5(1-=0,5)/379}} \leq -0.39$$

Região de aceitação

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

Outra forma de tomar a decisão



p-valor = $P(\hat{p} \leq 0,483 \mid p = 0,5)$

= 0,25 > **α** : decide por H_0

p-valor
- > α, decide por H_0
- < α, decide por H_a

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

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)

Hipóteses estatísticas de um teste (H_0, H_1)

É uma afirmação sobre os parâmetros da distribuição de probabilidade de uma ou mais populações.

$$H_0 \cap H_1 = \varnothing \quad e \quad H_0 \cup H_1 = \Theta$$

espaço paramétrico:
conjuntos de todos os possíveis valores do(s) parâmetro(s)

podem ser simples ou compostas
unilaterais (à esquerda, <, ou à direita, >) ou bilaterais (≠)

Teste de hipóteses

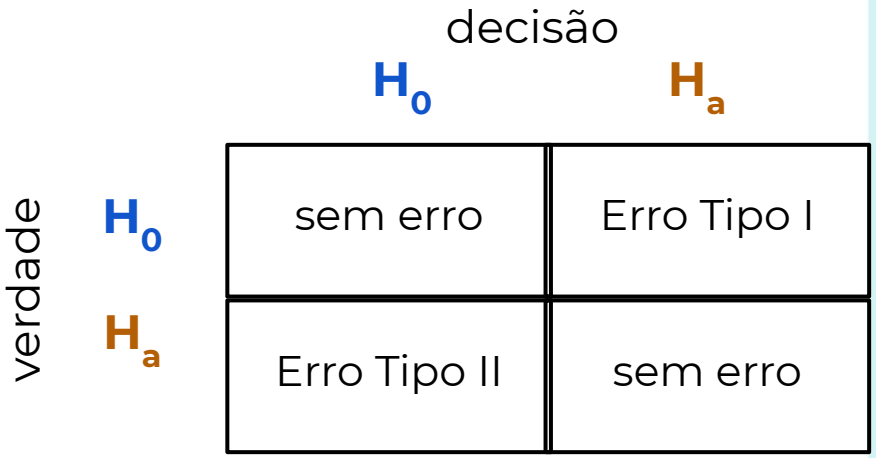
É um procedimento ou regra de decisão que nos leva a decidir por H_0 ou H_1 com base na amostra observada.

através da estatística do teste, sobre a qual se conhece a distribuição de probabilidades sob H_0

Região crítica ou de rejeição

Conjunto dos valores da estatística do teste em que a hipótese H_0 é rejeitada (i.e., decide-se por H_1).

A região complementar chama-se região de aceitação



nível de significância do teste

α = P(Erro Tipo I) = P(rejeitar H_0 | H_0 verdadeira) ↓

β = P(Erro Tipo II) = P(não rejeitar H_0 | H_0 falsa) = P(aceitar H_0 | H_0 falsa) ↓

$\pi = 1-\beta$ = P(rejeitar H_0 | H_0 falsa) ↑

poder do teste

Nível descritivo (ou valor p)

o menor valor do nível de significância para o qual H_0 é rejeitada.

p-valor
- $> \alpha$, decide por H_0
- $< \alpha$, decide por H_a

Críticas sobre o valor p (https://en.wikipedia.org/wiki/Misuse_of_p-values)

Misuse of *p*-values

From Wikipedia, the free encyclopedia

Misuse of *p*-values is common in scientific research and scientific education. *p*-values are often used or interpreted incorrectly; the American Statistical Association states that *p*-values can indicate how incompatible the data are with a specified statistical model.[1] From a Neyman–Pearson hypothesis testing approach to statistical inferences, the data obtained by comparing the *p*-value to a significance level will yield one of two results: either the null hypothesis is rejected (which however does not prove that the null hypothesis is *false*), or the null hypothesis *cannot* be rejected at that significance level (which however does not prove that the null hypothesis is *true*). From a Fisherian statistical testing approach to statistical inferences, a low *p*-value means *either* that the null hypothesis is true and a highly improbable event has occurred *or* that the null hypothesis is false.

Clarifications about *p*-values [edit]

The following list clarifies some issues that are commonly misunderstood regarding *p*-values: [1][2][3]

1. **The *p*-value is *not* the probability that the null hypothesis is true, or the probability that the alternative hypothesis is false.**[1] A *p*-value can indicate the degree of compatibility between a dataset and a particular hypothetical explanation (such as a null hypothesis). Specifically, the *p*-value can be taken as the prior probability of obtaining an effect that is at least as extreme as the observed effect, given that the null hypothesis is true. This should not be confused with the posterior probability that the null hypothesis is true given the observed effect (see prosecutor's fallacy). In fact, frequentist statistics does not attach probabilities to hypotheses.
2. **The *p*-value is *not* the probability that the observed effects were produced by random chance alone.**[1] The *p*-value is computed under the assumption that a certain model, usually the null hypothesis, is true. This means that the *p*-value is a statement about the relation of the data to that hypothesis.[1]
3. **The 0.05 significance level is merely a convention.**[2][4] The 0.05 significance level (alpha level) is often used as the boundary between a statistically significant and a statistically non-significant *p*-value. However, this does not imply that there is generally a scientific reason to consider results on opposite sides of any threshold as qualitatively different.[2][5]
4. **The *p*-value does not indicate the size or importance of the observed effect.**[1] A small *p*-value can be observed for an effect that is not meaningful or important. In fact, the larger the sample size, the smaller the minimum effect needed to produce a statistically significant *p*-value (see effect size). Visualizing effect sizes is a critical component of a data-analysis method called estimation statistics.

Multiple comparisons problem [edit]

Main article: Multiple comparisons problem

See also: P-hacking, Post hoc analysis, and Type I error

The multiple comparisons problem occurs when one considers a set of statistical inferences simultaneously[7] or infers a subset of parameters selected based on the observed values.[8] It is also known as the look-elsewhere effect. Errors in inference, including confidence intervals that fail to include their corresponding population parameters or hypothesis tests that incorrectly reject the null hypothesis, are more likely to occur when one considers the set as a whole. Several statistical techniques have been developed to prevent this from happening, allowing significance levels for single and multiple comparisons to be directly compared. These techniques generally require a higher significance threshold for individual comparisons, so as to compensate for the number of inferences being made.[*citation needed*]

The webcomic *xkcd* satirized misunderstandings of *p*-values by portraying scientists investigating the claim that eating jellybeans caused acne.[9][10][11][12] After failing to find a significant ($p < 0.05$) correlation between eating jellybeans and acne, the scientists investigate 20 different colors of jellybeans individually, without adjusting for multiple comparisons. They find one color (green) nominally associated with acne ($p < 0.05$). The results are then reported by a newspaper as indicating that green jellybeans are linked to acne at a 95% confidence level—as if green were the only color tested. In fact, if 20 independent tests are conducted at the 0.05 significance level and all null hypotheses are true, there is a 64.2% chance of obtaining at least one false positive and the expected number of false positives is 1 (i.e. 0.05 × 20).

In general, the family-wise error rate (FWER)—the probability of obtaining at least one false positive—increases with the number of tests performed. The FWER when all null hypotheses are true for *m* independent tests, each conducted at significance level α, is:[11]

$$\mathrm{FWER} = 1 - (1 - \alpha)^m$$

References [edit]

1. ^ a b c d e f g Wasserstein RL, Lazar NA (2016). "The ASA's statement on *p*-values: context, process, and purpose" (PDF). *The American Statistician*. **70** (2): 129–133. doi:10.1080/00031305.2016.1154108. S2CID 124084622.
2. ^ a b c Sterne JA, Davey Smith G (January 2001). "Sifting the evidence-what's wrong with significance tests?". *BMJ*. **322** (7280): 226–31. doi:10.1136/bmj.322.7280.226. PMC 1119478. PMID 11159626.
3. ^ Schervish MJ (1996). "*P* values: What they are and what they are not". *The American Statistician*. **50** (3): 203–206. doi:10.2307/2684655. JSTOR 2684655.
4. ^ Rafi Z, Greenland S (September 2020). "Semantic and cognitive tools to aid statistical science: replace confidence and significance by compatibility and surprise". *BMC Medical Research Methodology*. **20** (1): 244. doi:10.1186/s12874-020-01105-9. PMC 7528258. PMID 32998683.
5. ^ Amrhein V, Korner-Nievergelt F, Roth T (2017). "p > 0.05: significance thresholds and the crisis of unreplicable research". *PeerJ*. **5**: e3544. doi:10.7717/peerj.3544. PMC 5502092. PMID 28698825.
6. ^ Chaput, Brigitte; Girard, Jean-Claude; Henry, Michel (2011). "Frequentist Approach: Modelling and Simulation in Statistics and Probability Teaching". *Teaching Statistics in School Mathematics-Challenges for Teaching and Teacher Education*. New ICMI Study Series. **14**. pp. 85–95. doi:10.1007/978-94-007-1131-0_12. ISBN 978-94-007-1130-3.
7. ^ Miller RG (1981). *Simultaneous Statistical Inference* (2nd ed.). New York: Springer Verlag. ISBN 978-0-387-90548-8.
8. ^ Benjamini Y (December 2010). "Simultaneous and selective inference: Current successes and future challenges". *Biometrical Journal. Biometrische Zeitschrift*. **52** (6): 708–21. doi:10.1002/bimj.200900299. PMID 21154895.
9. ^ Munroe R (6 April 2011). "Significant". *xkcd*. Retrieved 2016-02-22.
10. ^ Colquhoun D (November 2014). "An investigation of the false discovery rate and the misinterpretation of p-values". *Royal Society Open Science*. **1** (3): 140216. arXiv:1407.5296. Bibcode:2014RSOS....140216C. doi:10.1098/rsos.140216. PMC 4448847. PMID 26064558.
11. ^ a b Reinhart A (2015). *Statistics Done Wrong: The Woefully Complete Guide*. No Starch Press. pp. 47–48. ISBN 978-1-59327-620-1.
12. ^ Barsalou M (2 June 2014). "Hypothesis testing and p values". *Minitab blog*. Retrieved 2016-02-22.