https://doi.org/10.4322/2357-9730.93649



# Teste de hipóteses: perguntas que você sempre quis fazer, mas nunca teve coragem

## *Hypothesis testing: questions you have always wanted to ask, but never had the courage to*

Vânia Naomi Hirakata[1], Aline Castello Branco Mancuso[1], Stela Maris de Jezus Castro[1,2]


## RESUMO

Dando continuidade aos artigos da série "Perguntas que você sempre quis fazer, mas nunca teve coragem", que tem como objetivo responder e sugerir referências para o melhor entendimento das principais dúvidas dos pesquisadores do Hospital de Clínicas de Porto Alegre sobre estatística, este segundo artigo se propõe a responder às principais dúvidas levantadas sobre Teste de Hipóteses. São discutidas questões referentes à metodologia de um teste de hipóteses na concepção clássica de Inferência Estatística, bem como tamanho de efeito, tipos de erros, valor de p e poder. Os conceitos são abordados numa linguagem acessível ao público leigo e diversas referências são sugeridas para os curiosos em relação ao tema.


***Palavras-chave:*** *Teste de hipóteses; inferência estatística; poder; tamanho de efeito; erro tipo I; erro tipo II; erro tipo III*


## ABSTRACT

Continuing the series of articles "Questions you have always wanted to ask, but never had the courage to", which aims to answer the most common questions of researchers at Hospital de Clínicas de Porto Alegre regarding statistics and to suggest references for a better understanding, this second article addresses the topic of hypothesis testing. The hypothesis testing method is discussed from a classical conception of statistical inference, including effect size, type of errors, p-value and power. The concepts are explained in plain language for lay readers and several references are suggested for those curious about the topic.


***Keywords:*** *Hypothesis testing; statistical inference; power; effect size; type I error; type II error; type III error*

Visando a publicação de artigos relevantes à necessidade dos pesquisadores no atual cenário estatístico, esta série "Perguntas que você sempre quis fazer, mas nunca teve coragem"[1] tem como objetivo responder e sugerir referências para o melhor entendimento das principais dúvidas levantadas em uma enquete encaminhada aos pesquisadores do HCPA – Hospital de Clínicas de Porto Alegre. Entre as 140 respostas analisadas na enquete sobre "quais assuntos você gostaria que fossem abordados na seção de Bioestatística?", além de sugestões de temas estatísticos, surgiram as mais variadas perguntas, das quais grande parte versavam sobre assuntos básicos de estatística ou já abordados na literatura. Diante desta realidade, o primeiro artigo da série respondeu algumas das principais questões levantadas sobre Estatística Descritiva[1]. Já este segundo artigo, tem como objetivo responder às principais dúvidas levantadas sobre Teste de Hipóteses, na concepção clássica de Inferência Estatística.

Clin Biomed Res. 2019;39(2):181-185

1 Unidade de Bioestatística, Grupo de Pesquisa e Pós-graduação (GPPG), Hospital de Clínicas de Porto Alegre (HCPA). Porto Alegre, RS, Brasil.

2 Departamento de Estatística, Instituto de Matemática, Universidade Federal do Rio Grande do Sul (UFRGS). Porto Alegre, RS, Brasil.

Autor correspondente:
Vânia Naomi Hirakata
l-bioestatistica@hcpa.edu.br
Hospital de Clínicas de Porto Alegre (HCPA)
Rua Ramiro Barcelos, 2350.
90035-903, Porto Alegre, RS, Brasil.



CC BY

## O QUE É UMA HIPÓTESE ESTATÍSTICA?

Em estatística, uma hipótese é uma afirmação sobre o valor de um parâmetro de determinada população, isto é, qualquer medida numérica calculada a partir de todos os indivíduos de uma população, por exemplo: média, proporção, etc. Uma afirmação do tipo "a concentração sérica do colesterol difere em crianças" não é uma hipótese estatística, pois não menciona o valor do parâmetro. Para tal, a afirmação deveria ser: "a média de concentração sérica do colesterol entre crianças de 7 a 12 anos é de 132 mg/dl"ou ainda "a diferença entre a média de concentração sérica do colesterol de crianças e adultos é nula[2].

## O QUE É UM TESTE DE HIPÓTESES?

O teste de hipóteses é um método de averiguação sobre a veracidade de uma afirmação, associado a um risco máximo de erro. Em outras palavras, por definição, um teste de hipóteses é uma regra de decisão para aceitar ou rejeitar uma hipótese, com base nas informações fornecidas pelos dados coletados em uma amostra e, por isso, envolve um risco de afirmar algo errado.

Devido à maneira como as análises são realizadas, cada teste de hipóteses inclui exatamente duas hipóteses sobre a população em estudo (nem mais, nem menos). Uma delas é chamada de hipótese nula ($H_0$), que é assumida como verdadeira até que se prove o contrário. A segunda é chamada de hipótese alternativa ($H_1$), que representa uma afirmação de que o parâmetro de interesse difere daquele definido na hipótese nula, de modo que as duas hipóteses sejam complementares.

Para desenvolver um teste de hipóteses existem certos passos que devem ser seguidos em 2 etapas: 1ª) ainda no planejamento do estudo; e 2ª) após a coleta de dados. Cabe salientar que além desta abordagem clássica de testes de hipóteses, também existe a abordagem Bayesiana[3], que se diferencia em vários aspectos.

### *1ª Etapa: no Planejamento do Estudo*

**Passo 1.1.** Estabelecer um objetivo e desfecho para o estudo

Uma questão importante a ser definida no início do planejamento de um estudo é o seu objetivo. Ele será a questão norteadora de todo delineamento e metodologia empregados. O objetivo de uma pesquisa quantitativa deve conter informações básicas e claras como: propósito (finalidade) e desfecho (variável em estudo)[4]. Para exemplificar, considere o desenvolvimento de um novo medicamento para controlar o nível de triglicerídeos, onde o objetivo (propósito) é comparar os níveis de triglicerídeos (desfecho) entre o grupo que utilizou um medicamento padrão e o grupo que utilizou o novo medicamento.

**Passo 1.2.** Formular as hipóteses nula ($H_0$) e alternativa ($H_1$)

Em estudos de comparação, por exemplo, é usual estabelecer como hipótese nula a inexistência de diferença entre os grupos. Como frequentemente é feita a comparação de um novo tratamento com um tratamento padrão, esta opção implica colocar o ônus da prova de efetividade no tratamento novo, uma opção conservadora, mas prudente[5]. Visto que prefere-se errar ao dizer que o novo medicamento não é melhor (quando na verdade é), do que errar ao dizer que um novo medicamento é melhor (quando na verdade não é).

A hipótese alternativa, por sua vez, será o complementar:

- Se a hipótese nula contempla igualdade ($H_0 : \mu_{tratamento\ A} = \mu_{tratamento\ B}$), então:
  $H_1 : \mu_{tratamento\ A} \neq \mu_{tratamento\ B}$ exemplo de teste bilateral).
- Se a hipótese nula contempla inferioridade ($H_0 : \mu_{tratamento\ A} \leq \mu_{tratamento\ B}$), então:
  $H_1 : \mu_{tratamento\ A} > \mu_{tratamento\ B}$ (exemplo de teste unilateral superior).
- Se a hipótese nula contempla superioridade ($H_0 : \mu_{tratamento\ A} \geq \mu_{tratamento\ B}$), então:
  $H_1 : \mu_{tratamento\ A} < \mu_{tratamento\ B}$ (exemplo de teste unilateral inferior).
- Se a hipótese nula for apenas um valor fixo ($H_0 : \mu_{colesterol} = 132$), então:
  $H_1 : \mu_{colesterol} \neq 132$

Tal que $\mu$ é a média da população/grupo em estudo. Perceba que a igualdade sempre é considerada na hipótese nula.

**Passo 1.3.** Decidir o nível de significância alfa (α)

O nível de significância é utilizado como um ponto de corte na probabilidade de se cometer um erro ao tomarmos a decisão estatística de rejeitar a hipótese nula (erro tipo I). Na área da saúde, o valor mais comum para o nível de significância é 0,05 (5%). Outros valores populares para nível de significância são 0,01 (1%) e 0,1 (10%).

### *2ª Etapa: Após a Coleta de Dados*

**Passo 2.1.** Definição do teste mais adequado

Neste ponto, os testes estatísticos já foram previamente listados no projeto de pesquisa, restando a decisão de uma abordagem paramétrica ou não, que deve ser feita baseada na distribuição da variável em estudo (do desfecho). Cabe salientar a importância da verificação das suposições do teste, se existirem.

**Passo 2.2.** Calcular a estatística de teste

A partir dos dados da amostra deve ser calculada a estatística de teste, sob a suposição de que a hipótese nula é verdadeira. A estatística de teste também incorpora a medida de erro padrão e suposições relacionadas à distribuição amostral do estimador do parâmetro que está sendo testado. Em suma, cada teste tem sua própria estatística, cujo valor calculado geralmente muda de uma amostra aleatória para outra.

**Passo 2.3.** Calcular o valor p

O valor p é a probabilidade da estatística de teste assumir o valor calculado no passo anterior ou um valor mais extremo, dado que a hipótese nula é verdadeira. Quando o valor p tende a ser pequeno (menor que o nível de significância adotado na etapa 1.3, pode-se fazer duas conjecturas: um evento que é extremamente raro pode ter ocorrido, ou a hipótese nula não deve ser verdadeira. Portanto, quanto menor o valor de p maior a evidência para se rejeitar a hipótese nula[5].

**Passo 2.4.** Decisão Estatística (rejeitar ou não a hipótese nula)

Ao final dos cálculos, toma-se uma das duas decisões estatísticas: rejeitar a hipótese nula ou não rejeitar a hipótese nula. Esta decisão deve ser tomada através da comparação do valor p com o nível de significância: quando o valor p for menor que o nível de significância estabelecido na etapa 1.3, toma-se a decisão de rejeitar a hipótese nula; caso o valor de p seja maior ou igual, toma-se a decisão de aceitar a hipótese nula.

**Passo 2.5.** Conclusão final

Uma vez que a decisão estatística foi tomada, devemos escrever o que ela representa em termos da hipótese testada, isto é, escrever uma conclusão para o teste. Esta conclusão deve ser tomada não apenas com os resultados estatísticos, mas também com base no conhecimento do pesquisador sobre o desfecho em estudo. Neste ponto, é importante ressaltar que "significância estatística" não equivale à "significância (ou relevância) clínica".

## O QUE É ESTATÍSTICA DE TESTE?

Uma estatística de teste, como visto na etapa 2.2, é utilizada em testes de hipóteses. Ela é calculada a partir dos dados da amostra e, por isso, também é considerada uma variável aleatória, visto que seu valor observado pode mudar de uma amostra aleatória para outra, mesmo que ambas tenham sido originadas da mesma população.

Em termos gerais, uma estatística de teste compara os dados amostrais com o que é esperado sob a suposição de que a hipótese nula é verdadeira. Em outras palavras, mede o grau de concordância entre uma amostra e a hipótese nula.

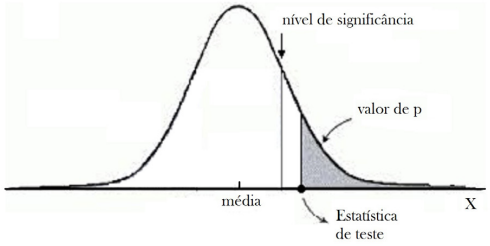


Figura 1: Valor de p em um teste unilateral.

## QUAL O SIGNIFICADO DO VALOR DE P?

Como já apresentado na etapa 2.3, o valor de p (*P-value*), também denominado de "nível descritivo do teste" ou "probabilidade de significância", está associado a um teste de hipóteses. Ele é calculado com base na estatística do teste de hipóteses em uso, na distribuição amostral da mesma e, ainda, no tipo de teste que está sendo feito (unilateral ou bilateral). A Figura 1 ilustra um exemplo unilateral.

A metodologia de cálculo do valor de p pode ser encontrado em diversos livros de introdução à estatística, tais como Soares e Siqueira[5], Kirkwood e Sterne[6] e Callegari-Jacques[2]. Recomenda-se, também, a leitura do trabalho de Boos & Stefanski[7] sobre a precisão e a reprodutibilidade do valor p.

A dificuldade na compreensão do significado do valor de p e seu uso incorreto acaba levando muitos pesquisadores a cometer graves equívocos na hora de discutir os resultados ou mesmo nas conclusões do trabalho[7-11]. Soares e Siqueira[5] já argumentavam que é inaceitável que resultados de dois estudos que apresentaram valores de p de 0.045 e 0.055, por exemplo, sejam interpretados de forma diferente para 5% de significância. Estes estudos deveriam levar a conclusões convergentes e não há conclusões diametralmente opostas.

## O QUE SIGNIFICA ERRO TIPO I, TIPO II E TIPO III?

Os significados de erro tipo I e erro tipo II estão vinculados ao conceito de teste de hipóteses. Quando se investiga alguma hipótese, como foi visto anteriormente, há duas possibilidades: ela realmente é verdadeira ou não. Visto que esta conclusão é baseada numa amostra, e a mesma não representa por completo a população, pode-se cometer um erro ao concluir que a hipótese nula é verdadeira quando na verdade não é, ou vice-versa (Tabela 1).

O erro tipo I também é conhecido como nível de significância (α), descrito nas etapas de um teste de hipóteses. Ele pode ocorrer quando, a partir de um resultado significativo, acredita-se que exista um efeito (diferença ou relação) na população em estudo, mas

Tabela 1: Erros tipo I e tipo II.

| Decisão | Realidade | |
|---|---|---|
| | **H0 verdadeiro** | **H0 falso** |
| **Aceita H0** | Decisão Correta | Erro Tipo II |
| **Rejeita H0** | Erro Tipo I | Decisão Correta |

na realidade este efeito não existe. Ou seja, rejeita-se a hipótese nula quando na realidade ela é verdadeira.

O oposto é o erro tipo II, conhecido também como erro-beta, que pode ocorrer quando não se encontra significância para provar um efeito que realmente existe na população. Ou seja, não se rejeita a hipótese nula quando na realidade ela é falsa. Este erro costuma ser fixado numa probabilidade de 0,2 ou menor e deve ser determinado ainda no planejamento do estudo, para o cálculo do tamanho da amostra.

Já o erro tipo III, menos conhecido e ainda pouco abordado, talvez por não ser um tipo de erro puramente estatístico, envolve basicamente o conceito do problema em questão. É o erro que cometemos quando rejeitamos a hipótese nula, mas aceitamos a hipótese alternativa errada. Ou seja, encontramos diferença ou relação significativa, mas por um motivo errado, ou ainda, a resposta certa para a pergunta errada. Nesse caso, a hipótese pode estar mal formulada ou incorreta. Por exemplo, testar a superioridade de um medicamento quando na verdade ele é inferior[12,13].

## QUAL O SIGNIFICADO DO PODER?

O poder de um teste estatístico é a probabilidade de se tomar uma decisão correta, rejeitar a hipótese nula se ela realmente for falsa. Ou seja, é a probabilidade complementar do erro tipo II (poder + probabilidade do erro tipo II = 1). Logo, à medida que o poder aumenta, as chances de um falso negativo (erro tipo II) ocorrer diminuem. Segundo O'Keefe[14], 3 aspectos podem influenciar o poder, considerando todos os demais parâmetros iguais:

1) O nível de significância: um teste possui maior poder se o nível de significância for 0.05 do que se for 0.01, por exemplo;
2) O tamanho da amostra: quanto maior a amostra, maior o poder;
3) O tamanho de efeito: quanto maior o tamanho de efeito, maior o poder do estudo;

## QUANDO CALCULAR O PODER?

De maneira geral, pode-se falar em poder prospectivo (determinado *a priori*), poder retrospectivo (a posteriori) e observado. A diferença básica entre prospectivo e retrospectivo está apenas no momento do estudo em que o poder é calculado, sendo ambos calculados a partir do nível de significância estabelecido e de dados já publicados de tamanho de efeito. O poder prospectivo é definido quando o estudo está sendo planejado e a coleta de dados ainda não iniciou (parte imprescindível do cálculo de tamanho de amostra quando o objetivo do estudo envolve testes de hipóteses) e o poder retrospectivo é aquele calculado após o estudo ter finalizado, mas ambos devem ser iguais[14]. Já o poder observado é calculado a partir dos dados coletados, mas, segundo O'Keefe[14], não se sabe por quê tal poder deve ser de interesse. Informações adicionais sobre o uso (ou abuso) dos diferentes tipos de poder também são discutidas em Thomas[15], Zumbo e Hubley[16] e Hoenig e Heisey[17].

## TAMANHO DE EFEITO: O QUE SIGNIFICA E COMO SE USA?

Em estatística, tamanho de efeito (*effect size*) é uma medida que quantifica a magnitude de um fenômeno de forma padronizada. O fato da medida ser padronizada nos permite comparar os efeitos sofridos em diferentes estudos que mediram diferentes variáveis ou usaram escalas distintas. Por exemplo, o tamanho de efeito de uma dieta baseado no IMC ($kg/m^2$) pode ser comparado a um efeito baseado no peso (kg), ou até mesmo no nível de triglicerídeos, colesterol, etc. E por isso vem sendo cada vez mais requisitado, pois é uma medida objetiva da importância de um efeito, independentemente do desfecho em estudo ou da significância estatística. Ou seja, é uma maneira de expressar a relevância (significância) clínica dos resultados encontrados.

Existem diversas medidas de tamanho de efeito descritas na literatura, pode-se citar: d de Cohen, que expressa diferença entre médias; razão de Odds e risco relativo para desfechos dicotômicos; coeficiente de correlação de Pearson as e coeficiente de determinação ($r^2$) para a relação entre duas variáveis quantitativas, entre muitas outras[18,19]. De forma geral, quanto maior o efeito maior a força de uma associação ou diferença. Cohen[18], Sullivan e Feinn[19] e Rosenthal[20] qualificam em seus artigos o que seria um efeito pequeno, médio ou grande.

O tamanho de efeito está intrinsecamente ligado a três outras propriedades estatísticas: o tamanho da amostra, o nível de significância (α) e o poder estatístico. Assim, dadas três dessas propriedades, podemos calcular a que falta[21]. Lindenau e Guimarães[22] demonstram como calcular alguns tamanhos de efeitos no SPSS. Mas softwares como o G-Power[23] e páginas da internet como Psychometrica[24], por exemplo, automatizam muitos dos cálculos a partir da entrada de algumas informações.

### *Conflitos de Interesse*

Os autores declaram não ter conflitos de interesse.

## REFERÊNCIAS


1. Mancuso ACB, Castro SMJ, Guimarães LSP, Leotti VB, Hirakata VN, Camey AS. Estatística descritiva: perguntas que você sempre quis fazer, mas nunca teve coragem. *Clin Biomed Res.* 2018;38(4):414-8. http://dx.doi.org/10.4322/2357-9730.89242.

2. Callegari-Jacques SM. Bioestatística: princípios e aplicações. 1. ed. Porto Alegre: Artmed; 2003.

3. Gelman A, Carlin JB, Stern HS, Rubin DB. Bayesian data analysis. 2nd ed. Boca Raton: Chapman & Hall/CRC Press; 2003.

4. Hulley SB, Cummings SR, Browner WS, Grady D, Hearst N, Newman TB. Delineando a pesquisa clínica: uma abordagem epidemiológica. Porto Alegre: Artmed; 2003.

5. Soares JF, Siqueira AL. Introdução à estatística médica. Belo Horizonte: Coopmed; 1999.

6. Kirkwood BR, Sterne JAC. Essential medical statistics. 2nd ed. Oxford: John Wiley & Sons; 2003.

7. Boos DD, Stefanski LA. P-value precision and reproducibility. *Am Stat*. 2011;65(4):213-21. http://dx.doi.org/10.1198/tas.2011.10129. PMid:22690019.

8. Goodman S. A dirty dozen: twelve p-value misconceptions. *Semin Hematol*. 2008;45(3):135-40. http://dx.doi.org/10.1053/j.seminhematol.2008.04.003. PMid:18582619.

9. Amrhein V, Greenland S, McShane B. Scientists rise up against statistical significance. *Nature*. 2019;567(7748):305-7. http://dx.doi.org/10.1038/d41586-019-00857-9. PMid:30894741.

10. Betensky RA. The p-value requires context, not a threshold. *Am Stat*. 2019;73(Suppl 1):115-7. http://dx.doi.org/10.1080/00031305.2018.1529624.

11. Krueger JI, Heck PR. Putting the p-value in its place. *Am Stat*. 2019;73(Suppl 1):122-8. http://dx.doi.org/10.1080/00031305.2018.1470033.

12. Robin ED, Lewiston NJ. Type 3 and type 4 errors in the statistical evaluation of clinical trials. *Chest*. 1990;98(2):463-5. http://dx.doi.org/10.1378/chest.98.2.463. PMid:2376179.

13. Stoltzfus J, Kaur P. Type I, II, and III statistical errors. *Int J Acad Med.* 2017;3(2):268. http://dx.doi.org/10.4103/IJAM.IJAM_92_17.

14. O'Keefe DJ. Brief report: post hoc power, observed power, a priori power, retrospective power, prospective power, achieved power: sorting out appropriate uses of statistical power analyses. *Commun Methods Meas*. 2007;1(4):291-9. http://dx.doi.org/10.1080/19312450701641375.

15. Thomas L. Retrospective power analysis. *Conserv Biol*. 1997;11(1):276-80. http://dx.doi.org/10.1046/j.1523-1739.1997.96102.x.

16. Zumbo BD, Hubley AM. A note on misconceptions concerning prospective and retrospective power. *J R Stat Soc Ser Stat.* 1998;47(2):385-8. http://dx.doi.org/10.1111/1467-9884.00139.

17. Hoenig JM, Heisey DM. The abuse of power. *Am Stat*. 2001;55(1):19-24. http://dx.doi.org/10.1198/000313001300339897.

18. Cohen J. A power primer. *Psychol Bull*. 1992;112(1):155-9. http://dx.doi.org/10.1037/0033-2909.112.1.155. PMid:19565683.

19. Sullivan GM, Feinn R. Using effect size: or why the p value is not enough. *J Grad Med Educ*. 2012;4(3):279-82. http://dx.doi.org/10.4300/JGME-D-12-00156.1. PMid:23997866.

20. Rosenthal JA. Qualitative descriptors of strength of association and effect size. *J Soc Serv Res*. 1996;21(4):37-59. http://dx.doi.org/10.1300/J079v21n04_02.

21. Field A. Descobrindo a statística usando o SPSS. 2. ed. Porto Alegre: Bookman; 2009.

22. Lindenau JDR, Guimarães LSP. Calculando o tamanho de efeito no SPSS. *Clin. Biomed. Res.* 2012;32(3):363-81. [citado 2019 Maio 12]. Disponível em: https://seer.ufrgs.br/hcpa/article/view/33160/22836

23. Faul F, Erdfelder E, Lang A-G, Buchner A. G*Power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behav Res Methods*. 2007;39(2):175-91. http://dx.doi.org/10.3758/BF03193146. PMid:17695343.

24. Lenhard W, Lenhard A. *Calculation of effect sizes*. Psychometrica; 2016. [citado 2019 Maio 12]. Disponível em: https://www.psychometrica.de/effect_size.html


*Recebido: 12 jun, 2019*
*Aceito: 18 jun, 2019*