

Capítulo 4

O Significado da Inferência Estatística

OBJECTIVO: Discutir o papel dos intervalos de confiança em confronto com os testes de significância da hipótese nula na decisão estatística; introduzir o conceito de magnitude do efeito (*effect size*); apresentar várias perspectivas de análise da potência do teste.

Palavras-chave: Análise do critério, Análise de sensibilidade, *d* de Cohen, *g* de Cohen, *g* de Hedges, Magnitude do efeito, *Meta-analysis*, Parâmetro de não centralidade, Potência *a priori*, Potência de compromisso, Potência *post hoc*, Significância estatística.

1. A doutrina da APA

A lógica dos testes estatísticos ou testes à significância da hipótese nula é simples, mas assenta numa dicotomia que, desde o primeiro contacto, surpreende como algo de dogmático: valores-*p* inferiores a um nível de significância arbitrário (ou, dito de outro modo, valores da estatística de teste superiores em valor absoluto a um determinado valor arbitrário) conduzem à rejeição da hipótese nula e, se assim não for, o resultado é inconclusivo. Mesmo um principiante não poderá deixar de se interrogar se será razoável tomar uma decisão tão drástica se o valor-*p* for aproximadamente igual ao nível de significância adoptado. E o que é que se deve entender por aproximadamente? Por esta e outras razões que se analisam neste capítulo, comprehende-se a necessidade da análise estatística focar a sua atenção, cada vez mais, na importância dos resultados para além da sua significância estatística.

Na convenção da *American Psychological Society* de 1996, um grupo de psicólogos e editores apresentou uma proposta radical de banir por completo a utilização dos testes de significância da hipótese nula da investigação científica. Tal proposta gerou enorme controvérsia entre os investigadores socio-comportamentais, tendo a APA criado a *Task Force for Statistical Inference*, integrando eminentes psicólogos conhecidos pela sua forte competência em estatística para estudar o assunto. Esta *Task Force*, embora clarificando que não apoiava qualquer acção que pudesse ser interpretada como banir o uso dos testes de significância da hipótese nula, criticou a sua má utilização e má compreensão e recomendou que, nas revistas da APA, esses testes fossem menos usados ou mesmo não admitidos. Diversos autores (Wilkinson *et al.* 1999; Nickerson, 2000) e editores de revistas científicas recomendam que os investigadores indiquem os valores-*p* dos testes de significância da hipótese nula (em vez de apenas indicarem nos seus relatórios que rejeitam ou não H_0 para um dado α), que os acompanhem de outras informações como a dimensão da amostra, a magnitude do efeito (*effect size*), a análise da potência e a análise de replicabilidade e que utilizem cada vez mais os intervalos de confiança.

Nesta secção, procura-se fazer um balanço dessa polémica e introduzir alguns conhecimentos complementares à introdução efectuada no Capítulo 3 sobre as alternativas ao papel da hipótese nula na investigação.

1.1. As críticas aos testes de significância

Muitas das críticas aos testes de significância da hipótese nula resultam dos investigadores lhes atribuírem qualidades que estes testes não têm ou de os utilizarem e/ou interpretarem incorrectamente.

Um erro frequente é a tendência para «aceitar» a H_0 , isto é, quando o resultado não é significativo, considerar a H_0 como verdadeira em vez de simplesmente considerar o teste inconclusivo como se viu no Capítulo 2.

Uma crítica importante aos testes de significância é que muitas teorias e métodos relevantes podem ser considerados falsos apenas por não conseguirem resultados suficientemente fortes, em virtude das amostras analisadas terem poucos participantes ou das medidas não serem exactas.

Também, os editores de revistas científicas rejeitaram, durante muitos anos, publicar estudos que relatassem testes inconclusivos por os considerarem pouco informativos. Este facto, conhecido como *viés editorial*, fez com que alguns conhecimentos fossem considerados verdadeiros durante muito tempo apenas porque alguns (poucos) estudos que os suportaram foram publicados, enquanto muitos outros em que a mesma H_0 não era rejeitada não o foram.

A valorização dos resultados estatisticamente significativos originou também que seja comum os autores utilizarem expressões incorrectas tais como «é quase significativo», quando relatam resultados estatisticamente não significativos, ou «altamente significativo», quando os resultados são estatisticamente significativos para um nível de alfa conservador (por exemplo 0,001).

Outra má interpretação frequente dos testes de significância é considerar-se «1-valor- p » como a probabilidade da decisão ser replicada em estudos futuros. Ora, a significância não prova a veracidade ou durabilidade da teoria. Thompson (2002) cita Cohen (1994) para lembrar que, para uma dada dimensão de amostra, os testes de significância apenas «estimam a probabilidade dos resultados de uma qualquer amostra se desviarem tanto ou mais que os resultados apresentados pela amostra actual dos resultados especificados na hipótese nula para a população». Assim, Thompson conclui: «estes testes não avaliam a probabilidade dos resultados da amostra descreverem a população; se estes testes o fizessem, suportariam que os resultados da amostra fossem replicáveis. Pelo contrário, os testes assumem que a hipótese nula descreve exactamente a população e depois testam a probabilidade para a amostra» (Thompson, 1996).

Refira-se,
tes de signifi
gadores qua

— a hipó
— as am
popula

1.2. Vant

Várias sã
confiança sej
ponto 2.1.1, o
testes bilatera
medida de pr
dos pelos test

Além destas
— são ex
permitti
— focam
significati
numa c
os inter
efeito e
— são pa
porque
medida
— viabiliza
produzir
uma va

(1) A *meta-analysis* é a combinação, integrar ou sintetizar

Refira-se, ainda, que, contrariamente ao que muitas vezes acontece, os testes de significância da hipótese nula não deveriam ser utilizados pelos investigadores quando

- a hipótese de efeito nulo for altamente improvável;
- as amostras utilizadas não forem seleccionadas aleatoriamente de uma população definível.

1.2. Vantagens dos intervalos de confiança

Várias são as razões invocadas para que se recomende que os intervalos de confiança sejam cada vez mais utilizados. Como já se referiu no Capítulo 3 no ponto 2.1.1, os intervalos de confiança contêm toda a informação chave dos testes bilaterais de significância da hipótese nula, podem ser utilizados como medida de precisão, assim como indicam todos os valores que não são rejeitados pelos testes bilaterais baseados numa certa amostra.

Além destas, outras vantagens são atribuídas aos intervalos de confiança:

- são expressos na mesma unidade de medida que a variável em estudo, permitindo uma fácil interpretação;
- focam a atenção na estimativa em vez da decisão, isto é, os testes de significância à H_0 apenas permitem conclusões sobre se um efeito vai numa direcção hipotética e nunca sobre a dimensão do efeito, enquanto os intervalos de confiança proporcionam uma ideia clara da grandeza do efeito e, ainda, da precisão da estimativa;
- são particularmente valiosos quando os resultados não são significativos porque, ao conhecer o intervalo, fica-se com uma ideia de em que medida o parâmetro se poderá afastar do efeito nulo;
- viabilizam a *meta-analysis*¹, que permite determinar se vários estudos produzem resultados significativamente diferentes ou qual o vigor de uma variável através de vários estudos.

(1) A *meta-analysis* consiste num conjunto de métodos quantitativos e gráficos utilizados para comparar ou sintetizar os resultados de uma vasta quantidade de estudos.

1.3. Razões para não abandonar os testes de significância

Há, no entanto, razões para não abandonar os testes de significância da hipótese nula. As mais frequentemente apontadas são:

- alguns métodos estatísticos avançados (análise multivariada e avaliação de ajustamento de modelos), não é praticável representar intervalos de confiança, mas é possível efectuar testes à significância da hipótese nula;
- tal como com os testes de significância, é igualmente possível, quando se constroem intervalos de confiança, cometer erros e interpretar mal os seus resultados, especialmente quando se é inexperiente.

Além disso, os testes estatísticos não são maus em si próprios, desde que usados como guias e indicadores e não como um meio de chegar a respostas definitivas (Huberty, 1987, p. 7).

Ao utilizar o teste à significância da hipótese nula para relatar os resultados, deve-se, no entanto, pelo menos, indicar o valor-*p* da estatística de teste, em vez de apenas indicar que se rejeita ou não a H_0 para determinado nível de significância. Por maioria de razão, nunca se deve, apenas simbolicamente, colocar uns asteriscos em frente da estatística de teste como surgem nos *outputs* de algum *software* estatístico.¹

Concluindo, o ensino dos testes de significância continua a justificar-se porque permanecem a metodologia mais utilizada na investigação, porque, mesmo que fossem totalmente banidos, são indispensáveis à compreensão da investigação feita durante décadas e ainda porque as suas alternativas são melhor compreendidas se forem conhecidos e utilizados adequadamente.

1.4. A interpretação de «significativo»

Muitos dos adversários dos testes de significância da hipótese nula realçam ainda que a razão porque se deve abandoná-los não reside apenas no facto destes terem tendência em ser mal utilizados e mal interpretados, mas que o

(1) Por vezes, os *outputs* do *software* de estatística e os autores usam * para indicar o nível de significância de 0,05, ** para 0,01 e *** para 0,001.

problema reside na própria essência desta metodologia. Na verdade, quando se rejeita (ou não) H_0 , diz-se que o resultado é (ou não) «significativo». Ora esta palavra é usualmente interpretada como sinónimo de «importante», embora a significância estatística tenha pouco ou nada a ver com a magnitude de uma diferença ou de uma relação.

Como se poderá constatar analisando as respectivas fórmulas, para uma mesma grandeza do efeito ou da relação, as estatísticas de teste tornam-se maiores à medida que aumenta a dimensão da amostra, viabilizando assim a rejeição da H_0 . Para um mesmo desvio entre a média da amostra e o respectivo valor médio da população, o valor- p diminui sensivelmente quando se tem uma amostra de 500 casos em vez de uma amostra de 50 casos, devido simplesmente ao aumento do número de graus de liberdade. Um coeficiente de correlação entre duas variáveis é estatisticamente significativo ao nível de 5% numa amostra de 10 casos se for superior a 0,632 mas, numa amostra de 100 casos, basta que seja superior a 0,192 e, numa amostra de 1000 casos, apenas superior a 0,062.

Na verdade, se a amostra for grande, um resultado com um efeito pequeno pode ser estatisticamente significativo e, nesse caso, pode-se admitir que existe um verdadeiro efeito diferente de zero, embora de pequena magnitude. Se, porém, a amostra for bastante pequena, um resultado com uma magnitude enorme pode não ter qualquer significância estatística e, nesse caso, o mais que se pode concluir é que a magnitude é grande, mas não se pode estar nada seguro de que um verdadeiro efeito realmente exista.

Deve-se assim realçar que «significância estatística» é uma expressão que não deve ser confundida com a acepção vulgar de «significância». Em suma, a *significância estatística* revela a força da evidência de que o efeito não é nulo, isto é, se há evidência suficiente que suporte a existência de um efeito. Não significa isto que a magnitude do efeito ou da relação seja grande. Torna-se portanto necessário recorrer a uma medida do efeito que seja independente da dimensão da amostra.

2. A magnitude do efeito¹

A 5^a edição do *Publication Manual of the American Psychological Association* (APA, 2001) assegura que, «para que o leitor compreenda completamente a importância das descobertas, é quase sempre necessário incluir algum indicador da magnitude do efeito» no relatório de um estudo. Dito de outro modo, ao apresentar os resultados de um estudo, o investigador deve fazer acompanhar o valor-*p* por uma medida da magnitude do efeito.

A *magnitude do efeito* (ES) é, numa acepção geral, a grandeza efectiva do resultado de uma investigação. As medidas da magnitude do efeito descrevem a grandeza de um resultado não afectada pela dimensão da amostra. Aumentando a dimensão de uma amostra, aumenta-se a significância do teste (o valor-*p* decresce), mas a magnitude do efeito permanece constante.

A magnitude do efeito proporciona informação que pode ser comparada com a de outros estudos e que pode ainda ser utilizada na acumulação de informação de estudos independentes, através da *meta-analysis*² que é, assim, uma alternativa à revisão da literatura «narrativa» tradicional.

O relato da magnitude do efeito pode ser feito através de vários índices. As duas famílias de medidas de magnitude do efeito mais importantes são a das diferenças que descrevem um acréscimo, um progresso, um benefício, etc. (por exemplo, são medidas desta família, o desvio da média de uma amostra em relação à média populacional expresso em unidades de desvio padrão ou a diferença entre as proporções de um atributo em dois grupos) e a das correlações que descrevem uma associação entre variáveis (por exemplo, são medidas desta família os coeficientes de correlação de Pearson, bisserial por pontos e ϕ)³.



Jacob Cohen

No caso
até agora, p
mula anterio

(1) Em inglês, *effect size*.

(2) Na *meta-analysis*, o ES é a medida comum que pode ser calculada para diferentes estudos e em seguida combinada numa análise global. A *meta-analysis* usa determinados estimadores do EF porque estes são independentes da dimensão da amostra. A estatística *t*, por exemplo, não poderia desempenhar este papel porque o seu valor é, como se viu, em parte, função da dimensão da amostra. Por exemplo, estudos com diferenças equivalentes entre as condições de tratamento experimental e de controlo podem ter estatísticas *t* muito diferentes se os estudos respectivos se

⁽³⁾ Outra família de medidas de magnitude do efeito é a dos *ratios* (por exemplo, a razão das *chances* (odds ratio) que mede a razão entre as probabilidades de dois resultados). Na literatura, é comum usar o termo *Odds* para se referir a Odds Ratio.

Cohen (1988) continua a ser a definição mais fraca, $d = 0$ é considerada naturalmente uma medida empírica (Cohen, 1988), sem atender

Apresentam-se nesta secção apenas as medidas de magnitude do efeito que se referem a uma única população, as quais devem ser associadas aos resultados dos testes estatísticos (Capítulo 2) e dos intervalos de confiança (Capítulo 3) que examinam a diferença entre um parâmetro e uma constante especificada na H_0 . Em cada um dos capítulos seguintes, apresentar-se-ão outras medidas de magnitude do efeito correspondentes a cada um dos *designs* de investigação que serão expostos.

2.1. Magnitude do efeito da média (d)



Jacob Cohen

Deve-se ao psicólogo Jacob Cohen (1988, 1992) o desenvolvimento do conceito de magnitude do efeito. Cohen propôs, no contexto da análise da potência, um conjunto de medidas da magnitude do efeito associadas aos testes estatísticos mais usuais (Cohen, J., 1988). No caso do teste *t*-Student, Cohen definiu uma medida do grau em que dois valores médios, μ_1 e μ_2 , diferem, expressa em unidades de desvio padrão populacional (σ). Este indicador é conhecido por *d de Cohen*, correspondendo a

$$d = \frac{\mu_1 - \mu_2}{\sigma} \quad (4.1)$$

No caso de a análise ser feita com um único grupo, tal como se considerou até agora, μ_2 é substituído pelo valor constante especificado na H_0 (C) e a fórmula anterior assume o seguinte aspecto

$$d = \frac{\mu - C}{\sigma} \quad (4.2)$$

Cohen (1988, 1992) propôs uma interpretação para o valor de *d* que continua a ser a mais utilizada pelos investigadores: $d = 0,20$ como uma magnitude fraca, $d = 0,50$ como moderada e $d = 0,80$ como forte. Estas balizas facilitam naturalmente a interpretação da magnitude do efeito e têm alguma consistência empírica (Cohen, 1988, pp. 24-27) mas não devem ser aplicadas com rigidez, sem atender ao contexto.

EXEMPLO 4.4. b)

Admita-se que se considera uma potência de 0,90 suficientemente elevada e que, por isso, se pretende diminuir a dimensão da amostra de modo a diminuir o tempo e os custos, mantendo o mesmo nível de significância, o teste unilateral e a magnitude do efeito ($d = 0,6(6)$). Qual o tamanho que deve ser definido para a amostra?

RESOLUÇÃO:

Recorrendo à tabela B5, com $\alpha = 0,01$ e $1 - \beta = 0,90$, tem-se um parâmetro de não centralidade de $\delta = 3,60$ e $n = \left(\frac{\delta}{d}\right)^2 = \left(\frac{3,60}{0,6(6)}\right)^2 = 29,16 \approx 30$.

EXEMPLO 4.4. c)

Pretende-se estimar a magnitude do efeito populacional mínima susceptível de ter sido detectada através de um estudo embrionário, em que se realizou um teste bilateral para prever uma eventualidade não expectável quanto à direcção da hipótese alternativa, com nível de significância 0,05, potência 0,80 e dimensão da amostra 25.

RESOLUÇÃO:

Sendo $\alpha = 0,05$ (teste bilateral) e $1 - \beta = 0,80$, tem-se $\delta = 2,80$.

De $\delta = d\sqrt{n}$, pode-se retirar $d = \frac{\delta}{\sqrt{n}}$, pelo que $d = \frac{2,80}{\sqrt{25}} = 0,56$.

Portanto, o teste só seria capaz de detectar uma magnitude do efeito superior a 0,56.

(4.14)

Actualmente, podem-se considerar cinco perspectivas principais de análise da potência: análise *a priori*, análise *post hoc*, análise da potência de compromisso, análise de sensibilidade e análise do critério (Faul *et al.*, 2007).

A análise da potência *a priori* (Cohen, 1988) proporciona o controlo da potência antes de o estudo ser efectuado e é recomendável quando se dispõe de amplos recursos para a recolha de dados. Esta análise permite determinar a dimensão da amostra em função do nível de significância α pré-especificado, do nível da potência desejado ($1 - \beta$) e da magnitude do efeito que se espera encontrar. O Exemplo 4.4.b ilustra este tipo de análise.

A análise da potência *post hoc* (Cohen, 1988) pode ser feita antes de o estudo ser efectuado, mas também permite examinar se um teste estatístico já efectuado tinha alguma chance de rejeitar uma H_0 falsa. A potência ($1 - \beta$) é determinada em função de α , da magnitude do efeito na população¹ e da dimensão da amostra estudada. O Exemplo 4.4.a ilustra este tipo de análise.

A análise da potência de compromisso é um conceito desenvolvido por Erdfelder (1984). Pode ser feita quer antes quer depois da recolha dos dados. Faz-se antes, quando se têm limitações para a dimensão da amostra (como é o caso de uma investigação sobre pacientes com uma doença rara ou simplesmente quando faltam recursos para recolher uma amostra maior), havendo que especificar *a priori* a dimensão da amostra viável. Faz-se depois da recolha dos dados, mas antes de os analisar, com o objectivo de encontrar um equilíbrio aceitável entre os riscos inerentes aos dois tipos de erro para a dimensão da amostra disponível. Quer α quer $1 - \beta$ são determinados em função da magnitude do efeito, da dimensão da amostra e de um *ratio* das probabilidades de erro $q = \beta/\alpha$. Uma discussão da potência de compromisso, tal como proposta por Cohen, é feita por Rosenthal e Rosnow (2008, 357-359). Dado que α é tipicamente definido como 0,05, Cohen (1965) recomendou que, na investigação comportamental, se adoptasse uma potência de pelo menos 0,80, o que significa um *ratio* de $q = \frac{\beta}{\alpha} = \frac{0,2}{0,05} = 4$. O Exemplo 4.5 ilustra este tipo de análise.

A análise populacionaliliar estudos jque um estespecificadodade permite do efeito que de análise.

A análise decisão, em tA análise do importante qude significâncde análise.

A Tabela

Tipo de
Análise da pot
Análise da pot
Análise da pot
Análise de ser
Análise do crit

(1) Note-se que se trata do parâmetro da magnitude do efeito e não da magnitude do efeito proporcionada pela chamada *análise da potência retrospectiva* que é estimada com base nos dados da amostra e usada para determinar a potência observada. A potência observada é algumas vezes apresentada nos *outputs* dos programas de *software* de análise estatística de dados. Ora, é improvável que a magnitude do efeito na amostra seja idêntica à magnitude do efeito da população respectiva (Zumbo & Hubley, 1998).

No âmbi especificar a poderá apare 2002, p. 228;

— estim na ba

A análise de sensibilidade utiliza-se para calcular a magnitude do efeito populacional em função de α , $1 - \beta$ e da dimensão da amostra. É útil para avaliar estudos já publicados, permitindo saber qual a magnitude do efeito mínima que um estudo, com uma determinada potência, dimensão da amostra e α especificado, é capaz de detectar. Noutra perspectiva, a análise de sensibilidade permite detectar se, com uma pequena amostra, a grandeza da magnitude do efeito que pode ser detectada é plausível. O Exemplo 4.4.c) ilustra este tipo de análise.

A análise do critério permite determinar α e, consequentemente, o critério de decisão, em função de $1 - \beta$, da magnitude do efeito e da dimensão da amostra. A análise do critério é uma alternativa à análise da potência *post hoc*. Pode ser importante quando, sendo α menos importante do que β , se pretende um nível de significância compatível com o β requerido. O Exemplo 4.6 ilustra este tipo de análise.

A Tabela 4.1 resume os cinco tipos de análise de potência.

Tabela 4.1

Tipo de análise da potência	O que se pretende determinar	O que se deve especificar
Análise da potência <i>a priori</i>	n	α , $1 - \beta$ e ES
Análise da potência <i>post hoc</i>	$1 - \beta$	α , ES e n
Análise da potência de compromisso	α e $1 - \beta$	ES, n e $\frac{\beta}{\alpha}$
Análise de sensibilidade	ES mínimo	α , $1 - \beta$ e n
Análise do critério	α	$1 - \beta$, ES e n

No âmbito da análise da potência, torna-se frequentemente necessário especificar a magnitude do efeito na população antes de realizar o estudo, o que poderá aparentemente parecer difícil. Várias formas têm sido propostas (Howell, 2002, p. 228; Rosenthal e Rosnow, 2008, p. 355) para o fazer:

- estimar o valor da magnitude do efeito, pelo menos aproximadamente, na base da investigação anterior ou de um estudo piloto;

- definir uma diferença considerada importante pelo investigador e usar um desvio padrão estimado através de outros dados;
- utilizar os valores de referência usados por Cohen (1988), por exemplo, para o d de Cohen, 0,20, 0,50 e 0,80.

A exiguidade de valores de α na Tabela B5 não permite realizar, em geral, a análise da potência de compromisso e a análise do critério.

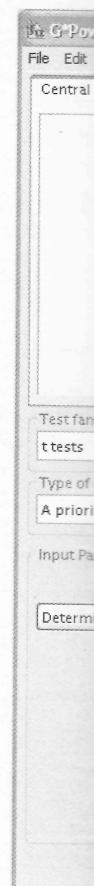
3.2. *G*Power*

O *G*Power 3* (Erdfelder, et al., 2007) é um *software* para PC que permite a análise da potência, segundo as cinco perspectivas descritas na secção anterior, para os testes estatísticos vulgarmente usados nas Ciências do Comportamento e ainda para todos os testes estatísticos que usam como referência as distribuições binomial, normal, qui-quadrado, *t*-Student ou *F* de Snedecor, desde que se introduza separadamente o parâmetro de não centralidade.

O cálculo da potência é feito através do *G*Power 3* em quatro etapas: 1) seleccionar o teste estatístico, 2) escolher uma das cinco perspectivas de análise, 3) inserir os *inputs* requeridos para a análise e 4) clicar no botão **Calculate**.

a) Seleccionar o teste estatístico

Pode-se seleccionar o teste estatístico de duas formas: através do menu *Test Family* na janela principal e em seguida no menu *Statistical test* que, entretanto, se adaptou automaticamente à família de testes seleccionada; ou através do menu expansível na barra de menus, onde se escolhe primeiro o tipo de parâmetro do teste pretendido (*correlação...*) e depois o *design* do estudo (*bisserial por pontos...*). A Figura 4.1 mostra que se escolheu a família dos testes para a correlação e regressão e o *design* bisserial por pontos.



b) Escolha

A escolha

Joaquim Pinto Coelho
Luísa Margarida Cunha
Inês Legatheaux Martins

Inferência Estatística

Com Utilização do SPSS e G*power

- Segundo as recomendações da APA
- Magnitude do efeito numa investigação
- Intervalos de confiança
- Testes estatísticos
- Análise da potência
- Determinação da dimensão da amostra



EDIÇÕES SÍLABO