Que confiança podemos ter nas conclusões estatísticas que apresentamos ?

Manuela Azevedo
Instituto Português de Investigação Marítima
Teresa Garcia-Marques
Instituto Superior de Psicologia Aplicada

Resumo. - Os testes estatísticos são um instrumento utilizado em investigação com vista a decidir/concluir sobre a presença/ausência de um efeito. Este artigo aborda a questão da potência dos testes estatísticos nas suas implicações para a validade das conclusões da análise de uma investigação. São também discutidos os dois tipos de erros associados à abordagem inferencial e as suas relações com a hipótese que rege a investigação (se coincidente com a hipótese nula, Ho, ou não). Paralelamente são discutidos assuntos como significância estatística, significância pragmática e representatividade.

Um projecto de investigação visa a recolha de dados informativos sobre a realidade em estudo. O modo como se recolhem esses dados é determinante da sua informatividade, fornecendo critérios de avaliação do projecto que se associam à validade das conclusões que este proporciona, usualmente (na tradição de Campbell, 1957) distinguidos em critérios de validade interna, externa e de validade das suas operacionalizações (validade de construto). Estes critérios, felizmente, tendem a estar presentes na concepção da maioria dos nossos projectos de estudo. No entanto, frequentemente tendemos a negligenciar um outro critério que define igualmente a validade das conclusões dos nossos estudos: a sua validade estatística ou validade relativa à conclusão da existência de uma covariação nos nossos dados (Cook & Campbell, 1979). Esta validade associa-se a um conjunto de questões das quais destacamos: (1) a adequação das técnicas estatísticas utilizadas para inferir a existência de covariação (utilização do modelo estatístico apropriado à natureza dos nossos dados - validade dos seus pressupostos); (2) a sensibilidade do nosso estudo e das técnicas estatísticas para a detecção de covariação; (3) a probabilidade, associada ao procedimento estatístico, de se inferir erradamente a presença de covariação. A preocupação dos investigadores, regra geral, recai quase unicamente sobre esta última questão (embora nem sempre de uma forma exaustiva, ver Garcia-Marques & Azevedo, 1994). A adequação do procedimento estatístico é quase sempre tratada de uma forma muito superficial - a que atende meramente à escala de medida das variáveis em estudo - embora se observe por vezes preocupação em testar os pressupostos subjacentes à construção do teste a realizar. A segunda das questões levantadas relaciona-se com a potência dos testes realizados para concluir pela presença ou ausência de uma relação entre as variáveis em estudo, e sobre ela se centrará este artigo.

Ao apresentar e discutir a questão da potência de teste, abordaremos as estratégias à disposição do investigador, que procuram garantir um justo julgamento da sua(s) hipótese(s). Paralelamente, discutiremos a tendência errada de se interpretar um resultado não significativo (p> nível de significância estabelecido) como sinónimo de aceitação da hipótese nula e evidenciaremos a necessidade do investigador não se preocupar com os erros de Tipo I, em detrimento dos erros de Tipo II, quando postula a hipótese nula como verdadeira.

Em 1962, Jacob Cohen estimula a discussão das questões relativas à potência dos testes estatísticos, ao publicar uma análise da potência dos testes associados a um conjunto de artigos de investigação em Psicologia. Os seus argumentos ganharam peso ao demonstrarem que algumas conclusões de estudos no campo da Psicologia haviam sido erradamente induzidas, em consequência da fraca potência dos testes estatísticos realizados[1]. A realidade, em 1962, era a de que, em média, para efeitos de magnitude moderada, os testes estatísticos tinham uma potência de 48% (pelo que a probabilidade destes testes detectarem um efeito, que se sabe verdadeiro, era idêntica à probabilidade associada ao acaso). Após 24 anos, uma análise semelhante à de Cohen é levada a cabo por Sedlmeier e Gigerenzer (1989), e os resultados não foram melhores, sendo aquela média de 50%. Desta forma, os efeitos significativos apontados naqueles estudos teriam, em média, 50% de hipóteses de serem detectados por outros investigadores que pretendessem replicar as experiências. Nenhuma análise semelhante foi realizada sobre os estudos portugueses do campo da Psicologia, pelo que, numa perspectiva optimista, os testes terão, quando muito, uma potência semelhante às análises referidas. Que confiança podemos ter, então, nas conclusões estatísticas que apresentamos, se trabalhamos com instrumentos de tão fraca potência? A situação é deveras preocupante? Que medidas podemos tomar com vista a mudar a situação?

A situação é apenas parcialmente preocupante, visto que, embora não possamos ter confiança em conclusões que põem em causa a presença de covariação, ela não afecta a validade das inferências que lhe dão suporte. A conclusão sobre a presença de um efeito apenas pode estar errada se esse efeito não existir na população e este erro é independente da potência de um teste para detectar a sua presença. Por outro lado, a gravidade da situação pode ser atenuada por algumas iniciativas por parte do investigador, como veremos neste artigo.

Alguns conceitos básicos.

Com vista a tornar explícita a noção de potência de teste, recordemos alguns conceitos básicos que lhe estão associados.

[1] Este problema estende-se a outras áreas científicas como a Biologia (Peterman, 1990), onde é igualmente vulgar referir a não rejeição da hipótese nula como sinónimo de sua aceitação.

Não é de certo estranha ao leitor a Tabela 1, onde se ilustra o tipo de erros passíveis de serem cometidos quando se tomam decisões sobre a validade da hipótese especificada pela hipótese nula (Ho), associada a um dado teste estatístico . O *erro Tipo I* ocorre quando rejeitamos Ho, sendo esta verdadeira. A probabilidade de cometer este tipo de erro define o *nível de significância* de uma conclusão, que é especificado *a priori* pelo investigador. Por outro lado, se Ho for falsa e a tomarmos como verdadeira, i.e., não a rejeitarmos, cometeremos um *erro Tipo II*. As probabilidades de obtenção destas conclusões falsas são os riscos de 1ª e de 2ª espécie, também designados por a e ß, respectivamente (ver tab. 1).

Tabela 1.
Tipos de erro que se podem cometer ao testar estatísticamente uma determinada hipótese nula (Ho).

	Estado da natureza	
	Ho verdadeira	Ho falsa
Decisão Rejeitar Ho	erro Tipo I (α)	
Não rejeitar Ho		erro Tipo II (ß)

A decisão estatística será correcta quando: (i) se rejeitar Ho, sendo Ho falsa ou (ii) não se rejeitar Ho, sendo Ho verdadeira. Deste modo, 1-α define o *coeficiente de confiança* do teste realizado (a probabilidade de não se estar a cometer, por acaso, um erro de Tipo I) e 1-ß, sendo a probabilidade de não estarmos a cometer, por acaso, um erro de Tipo II , define a *potência* do teste. A relação entre a probabilidade de rejeição e o grau de falsidade da Ho constitui a *função potência* de teste.

Pretendendo reduzir a probabilidade de retirar conclusões erradas do seu estudo, o investigador só considera um efeito como presente, caso a probabilidade de o observar, por mero acaso, seja realmente muito pequena. Estando directamente sobre o seu controlo o estabelecimento do nível de significância, o investigador estabelece um reduzido risco de erro Tipo I ao definir, regra geral, α=5% ou 1%. O controlo do erro Tipo II e da potência do teste, não é de modo algum tão directo. A potência de um teste é função da magnitude do efeito sob análise, do nível de significância escolhido pelo investigador para a análise dos seus dados, e da precisão/fidelidade dos dados recolhidos. Esta precisão reporta-se directamente ao termo de erro da mensuração do efeito a estudar, definido pela razão de s^2/n (erro padrão: em que s^2 representa a variância empírica e n a dimensão da amostra). A potência de um teste relaciona-se, ainda, com a especificidade do teste estatístico a levar a cabo (Cohen, 1988).

Atendamos separadamente a cada um dos três parâmetros que definem a potência de um teste estatístico (Cohen, 1988): (i) o nível de significância, a, (ii) a magnitude do efeito, *ME*, e (iii) a dimensão da amostra, *n*, relativa à variabilidade observada nos dados, s^2.

Nível de Significância, α.

A potência de um teste está positivamente relacionada com alfa, dado esta probabilidade determinar directamente a área da distribuição de amostragem que define o erro Tipo II.
A prática vulgar de seleccionar nos testes de hipóteses um pequeno, seguindo o raciocínio de que «quanto mais pequeno, melhor», resulta, regra geral, em valores de potência (1-ß) relativamente pequenos e, portanto, em elevada probabilidade de cometer um erro Tipo II (ß). Vejamos o caso em que o investigador, ao seleccionar um α=0.001, obtém uma potência de teste de 10% (i.e.,1-ß=0.1). Note-se que, neste caso, a probabilidade de cometer pelo menos um erro Tipo II é de 90%, pois ß=0.9, pelo que seria sensato proceder a uma revisão do plano experimental, incluindo a revisão do nível de α de modo a aumentar a potência do teste. Mas, caso o investigador queira prosseguir com as análises, então deve ter consciência do ratio $^{\text{erro Tipo II}}/_{\text{erro Tipo I}}$ que vem a ser neste caso ß/α = 0.9/0.001 (um ratio 900:1), ou seja, implicitamente o investigador acredita que rejeitar erradamente Ho é 900 vezes mais «sério» do que não rejeitar Ho erradamente (Cohen, 1988).

Magnitude do Efeito, *ME*.

O fenómeno ou efeito que se pretende estudar traduz-se em ausência de covariação nos dados observados (se a hipótese nula for verdadeira) ou em presença dessa covariação (se a hipótese nula for falsa). No entanto, a realidade não assume o mesmo carácter dicotómico de presença e ausência, pelo que se o efeito estiver presente na população que se pretende estudar, ele pode manifestar-se com diferentes magnitudes. Ora, um teste pode ser suficientemente potente para detectar um efeito de elevada magnitude e não detectar um efeito de magnitude mais reduzida. Efeitos de elevada magnitude traduzem-se em distribuições de amostragem com médias bastante distanciadas da média da distribuição associada ao efeito nulo, pelo que a probabilidade de erro Tipo II é reduzida proporcionalmente ao aumento dessa distância.
Tomemos o exemplo hipotético em que conhecíamos a percentagem correcta de indivíduos do sexo masculino de uma dada população, e que esta era de 52%. Um investigador que não partilhasse do nosso conhecimento, realiza um estudo com o objectivo de determinar a composição sexual daquela população. Recolhe uma amostra que lhe permita testar Ho:%M=%F *vs* H1: %M≠%F. Ao estabelecer aquela hipótese nula, o investigador implicitamente está a «dizer» que a magnitude do efeito da variável sexo é zero. Mas, à luz do que sabemos, a magnitude do efeito é de dois pontos percentuais (medida como a percentagem de afastamento da percentagem definida em Ho), pelo que Ho deveria ser rejeitada. Mas Ho pode, no entanto, não ser rejeitada apenas por falta de sensibilidade do instrumento estatístico utilizado. A probabilidade de interpretar um efeito de magnitude elevada como ausente, será tanto maior quanto menor for a potência de teste.
Qualquer que seja a escala de medida em que se defina a magnitude do efeito

(ME), este pode ser tratado como um parâmetro que assume o valor zero quando Ho é verdadeira e qualquer outro valor diferente de zero quando Ho é falsa. Deste modo, a ME mede a discrepância entre a hipótese nula, Ho, e a hipótese alternativa, H1, (Cohen, 1992).

Dimensão da Amostra (n) e Variabilidade Observada (s^2).

A dimensão da(s) amostra(s), n, está estreitamente relacionada com a precisão a obter na estimação de parâmetros da população de dimensão N, sendo determinante da magnitude do erro padrão (s^2/n) associado a essa estimação. No caso extremo em que n=N, os parâmetros da população são estimados com um coeficiente de confiança de 100%, pelo que não existe um problema de estimação propriamente dito, visto se reduzir a zero a probabilidade de cometer erros de inferência. Tal, no entanto, pressupõe uma total precisão do instrumento de medida usado (ausência do *erro aleatório de medição*) e o controlo de qualquer outra fonte de variabilidade associada à recolha dos dados. O planeamento cuidadoso da experiência, com vista à eliminação de fontes de variabilidade irrelevantes para a avaliação do fenómeno em estudo, reduzindo s^2 e, portanto, o erro padrão aumenta a potência do(s) teste(s).

Na prática, o investigador raramente trabalha com a população que pretende estudar, mas com uma amostra reduzida de elementos dessa população, pelo que a dimensão dessa amostra é determinante da precisão associada aos seus dados. A relação de n com a potência de teste é intuitiva: quanto maior a precisão dos resultados obtidos com a amostra, maior a probabilidade de detectar o fenómeno ou efeito explicitado em H1.

O Modelo Estatístico e a Potência do Teste Associado.

Dentro das técnicas estatísticas criadas com vista à detecção de covariação, existem umas mais potentes que outras. Em particular, a abordagem paramétrica apresenta-se mais sensível à detecção de efeitos do que a não-paramétrica, regra geral mais conservadora. A potência de teste é igualmente afectada pelo grau de ajustamento dos dados aos pressupostos do teste estatístico a realizar (Peterman, 1990).

A direcção da região de rejeição de um teste, que define um teste como uni- ou bilateral, também interfere na potência do teste. Quando Ho pode ser rejeitada em ambas as direcções (teste bilateral), o teste resultante tem menor potência que um teste unilateral com o mesmo α, desde que o resultado obtido com aquela amostra vá na direcção prevista (Cohen, 1988). Assim, o teste de Ho relativamente à hipótese alternativa $m_A > m_B$, a um nível de alfa de 5%, terá uma potência equivalente ao teste bilateral com α= 1%. Isto, caso m_A seja realmente maior que m_B, pois se $m_A < m_B$, o teste não tem potência, uma vez que esta afirmação não é contemplada na hipótese alternativa, logo inadmissível. Deste modo, estudos que definam claramente as suas hipóteses experimentais, podem e devem definir as hipóteses alternativas dos seus testes como unilaterais, deixando a hipótese de efeito não direccionado apenas a

casos específicos como, por exemplo, o caso dos estudos exploratórios.

Estimação de parâmetros associados à potência de teste.

As expressões que permitem determinar a potência de um teste englobam os parâmetros já referidos, mas são específicas ao método estatístico particular que vai ser utilizado (teste-t, ANOVA, etc.). Essas expressões podem ser resolvidas relativamente a qualquer um daqueles parâmetros, desde que se estabeleçam os restantes como fixos *a priori*. Podemos assim, (1) determinar qual a potência de um teste associado a um dado α , n e *ME*, (2) determinar a dimensão da amostra associada a uma dada potência de teste, com α e *ME* pré-determinados, (3) determinar a magnitude de efeito, passível de ser detectada por um teste, com uma determinada potência, com uma amostra de dimensão n e um dado α, e, por último, (4) determinar que risco de erro Tipo I poderíamos cometer, caso pretendêssemos uma potência de teste x, com aquele n e para aquela *ME*.

Jacob Cohen (1988), pretendendo facilitar todo este tipo de análises, concebe tabelas dos seus valores relativas a algumas das situações mais vulgares na investigação em Psicologia. Requerendo do utilizador poucos conhecimentos matemáticos para a análise da potência de teste, Cohen oferece igualmente um conjunto de exemplos ilustrativos que esclarecem dúvidas eventuais. A sua extensão e profundidade é, no entanto, muitas vezes dissuasora de uma leitura atenta, por um leitor menos entusiasta com a preocupação com a potência dos seus testes. Para estes recomendamos o trabalho de Kraemer & Thiemann (1987), que publica uma tabela única para a estimação de valores dos parâmetros, associados a diferentes potências de teste de diferentes testes estatísticos, e que as autoras designaram de *Master Table*.

Para o leitor ainda menos motivado para o assunto, Cohen (1992) escreve um artigo apresentando uma simples regra de algibeira para o cálculo da dimensão da amostra, necessária a uma determinada potência de teste para o recurso a análises estatísticas mais comuns.

Infelizmente, os programas estatísticos mais comumente utilizados em investigação[2], carecem de algoritmos que permitam a estimação de parâmetros associados a diferentes potências de teste. Em Goldstein (1989) pode ler-se uma revisão comparativa dos programas disponíveis para computadores MS/PC-DOS, elaborados para a determinação da dimensão da amostra e da potência de teste.

Em termos práticos, não se encontra directamente sob o controlo do investigador o estabelecimento quer da magnitude do efeito (que define o objecto de estudo do investigador) quer do erro aleatório de mensuração desse efeito (que apenas pode ser considerado como reduzido ao máximo, por um cuidadoso planeamento do estudo que reduza as

[2] O programa STATISTICA apresenta no seu módulo de Process Analysis a possibilidade de estudo da potência de teste (bem como o recurso a técnicas de análise sequencial), para comparação de médias de duas populações.

fontes de variabilidade estranhas que se lhes associem). Além disso, enquanto os restantes parâmetros associados à potência de teste se apresentam numa escala absoluta pré-estabelecida, a magnitude de efeito traduz-se em escalas arbitrárias. Há, assim, necessidade de uniformizar as unidades a atribuir à *ME*, de modo a consultar qualquer das tabelas de potência de teste ou mesmo os gráficos de Pearson e Hartley (1951), que surgiram como a primeira estratégia de cálculo. Os índices usados combinam o efeito e a sua variabilidade, e dependem do tipo de análise estatística que se pretende realizar. Reportando o leitor a alguma das obras acima referidas, ilustraremos apenas alguns dos casos, nomeadamente i) o da comparação das médias de duas populações (teste-*t* para amostras não emparelhadas), ii) uma análise de variância simples com *k* níveis ou grupos e iii) o teste de comparação de coeficientes de correlação. Em todos estes casos, considera-se que não houve violação dos pressupostos inerentes aos métodos estatísticos em questão e usa-se a notação de Cohen (1988) para os respectivos índices de magnitude de efeito.

i) Considere-se o teste-*t* de comparação das médias de duas amostras independentes, sob a hipótese de que a diferença das médias das respectivas populações é zero. Assim, seja:

$$Ho: m_A = m_B \; vs \; H1: m_A \neq m_B$$
$$nA = nB$$
$$\sigma^2_A = \sigma^2_B = \sigma^2$$

O índice *d*, da magnitude de efeito associado a esta análise, é dado pela diferença absoluta das médias das populações padronizada pelo desvio padrão das populações, σ (por sua vez estimado pelo desvio padrão empírico):

$$d = \frac{m_A - m_B}{\sigma}$$

Se o teste é unilateral, aquela diferença deixa de ser absoluta para tomar valores positivos ou negativos consoante o sentido da diferença especificado na hipótese unilateral.

Para se ter uma ideia da ordem de grandeza dos níveis de *d*, Cohen propõe uma escala funcional para este índice dada por

d = 0.2 se o efeito é pequeno
d = 0.5 se o efeito é médio
d = 0.8 se o efeito é grande

ii) No caso da ANOVA, com um factor (modelo de efeitos fixos) se Ho é verdadeira, tanto a média da diferença quadrática entre grupos, MSA, como o quadrado médio residual, MSe, obtidos do modelo são uma estimativa da variância comum às *k* populações, σ^2, considerando-se apenas uma popula-ção. Se as médias não são iguais, então MSA será maior que MSe (Scheffé, 1959) e a razão entre aqueles quadrados médios segue uma distribuição F[glA, gle] (em que *glA* e *gle* representam o número de graus de liberdade associados à fonte de variação devida ao factor A e ao erro residual, respectivamente). No entanto, se Ho é falsa, aquela razão segue uma distribuição F não-central com parâmetros *glA*, *gle* e λ (designado parâmetro de não-centralidade). Como a potência de teste se refere à probabilidade de detectar se a hipótese nula é falsa, o cálculo da potência no caso da análise de variância depende desta distribuição F não-central.

No caso da análise de variância (modelo de efeitos fixos) a um factor de k níveis e igual número de observação por célula, o índice da magnitude do efeito, f, virá dado por:

$$f = \frac{\sqrt{\frac{\sum_{i=1}^{k}(m_i - m)^2}{k}}}{\sqrt{MSe}} = \frac{\sqrt{\text{variância do factor}}}{\sqrt{\text{variância residual}}}$$

em que m_i (média do nível i: $i=1,...,k$); m (média global) e MSe (variância residual do modelo ANOVA). Este índice representa também o desvio padrão das médias padronizado pelo desvio aleatório, sendo por tal independente da escala de medida, ou seja, *scale-free* (Cohen, 1992).

Um outro índice vulgarmente empregue para quantificar a magnitude do efeito é o índice designado por omega quadrado, ω^2 (Keppel, 1991). A relação entre ω^2 e f é dada pela expressão $\omega^2=f^2/(1+f^2)$, que representa uma medida da dispersão dos m_i (Cohen, 1988). Quando as médias das populações são todas iguais $f^2=0$ e, portanto $\omega^2=0$, indicando que nenhuma parte da variância total se deve a diferenças naquelas populações. Estes índices reflectem a proporção da variabilidade total observada na experiência que é explicada pela variabilidade entre os diferentes tratamentos.

Os níveis de grandeza convencionais atribuídos a estes índices, com o objectivo prático de se ter uma perspectiva de interpretação dos seus valores, têm a seguinte correspondência:

$f = 0.10$; $\omega^2 = 0.01$ se o efeito é pequeno
$f = 0.25$; $\omega^2 = 0.06$ se o efeito é médio
$f = 0.40$; $\omega^2 = 0.14$ se o efeito é grande

iii) Num teste de comparação de dois coeficientes de correlação em amostras independentes, a determinação do índice da magnitude do efeito, tendo em consideração a magnitude individual de cada coeficiente de correlação (r_1 e r_2: coeficientes de correlação de Pearson) que é obtida usando a transformação de Fisher ($z_1 = 0.5 \ln(1+r_1/1-r_1)$; $z_2 = 0.5 \ln(1+r_2/1-r_2)$), corresponde à magnitude da sua diferença. Assim, o índice da magnitude do efeito, q, para o teste i) bilateral ou ii) unilateral, vem dado por:

i) $q = | z_1 - z_2 |$
ii) $q = z_1 - z_2$ ou $q = z_2 - z_1$

A escala de Cohen para as diferenças entre coeficientes de correlação corresponde aos seguintes níveis de q:

$q = 0.1$ se o efeito é pequeno
$q = 0.3$ se o efeito é médio
$q = 0.5$ se o efeito é grande

Para ilustrar esta escala de magnitude de diferença indicam-se alguns valores para os coeficientes de correlação, r. Assim, uma «pequena» diferença entre coeficientes de correlação (q=0.1) corresponderá, por exemplo, a valores de pares de r como (0.00, 0.10); (0.40,0.48); (0.80,0.83) ou ainda (0.95,0.96). A uma magnitude de efeito médio (q=0.3), aos pares (0.00, 0.29); (0.40,0.62); (0.80,0.89) ou ainda (0.95,0.97) e para magnitudes de efeito grande (q=0.5), aos pares (0.00,

0.46); (0.40, 0.73); (0.80,0.92) e finalmente (0.95,0.98). Estes exemplos ilustram igualmente por que razão a mera diferença entre os coeficientes de correlação não é uma medida adequada da magnitude deste efeito.

Estes índices, como já referimos, são medidas de magnitude de efeito padronizadas, pelo que incorporam a estimativa da variância residual, s^2 (daí a sua ausência nas tabelas apresentadas por Cohen, 1988).

Vejamos um exemplo ilustrativo da relação entre os parâmetros referidos, considerando um estudo que envolve a comparação de dois valores médios por um teste-*t*. A Tabela 2, tendo em consideração a escala de três valores *d* de magnitude de efeito, definida por Cohen para aquele teste, estabelece a relação entre a potência de teste e a dimensão das amostras.

Tabela 2.
Dimensão da amostra, *n*, em função da potência de teste, da magnitude de efeito e do nível de significância, α, para a realização de um teste-*t* bilateral.

Potência	*Magnitude do efeito (d)*		
(1-ß)	0.2	0.5	0.8
α=0.05			
0.25	84	14	6
0.50	193	32	13
0.75	348	57	23
0.85	450	73	29
0.95	651	105	42
α=0.01			
0.25	183	31	13
0.50	333	55	22
0.75	528	86	35
0.85	654	106	43
0.95	892	144	57

Façamos uma primeira leitura da Tabela 2. Fixemos o nível de α em 0.05. Para uma magnitude de efeito pequena (d=0.2) e para uma potência de 0.75, a dimensão requerida para cada amostra é n=348 observações. Se a magnitude do efeito for média (d=0.5), então para a mesma potência de teste o número de observações em cada amostra vem *n*=57 elementos. Este valor diminui para 23 observações por amostra se a magnitude do efeito for grande (*d*=0.8). Esta tendência decrescente de *n* com o aumento da ME é notória em cada nível de potência de teste. Verificamos assim que, quanto maior a ME definida, maior a capacidade de detectar o efeito com uma amostra de menor dimensão.

Numa segunda leitura daquela Tabela, fixemos também a magnitude do efeito em d=0.2. A dimensão de cada amostra varia entre 84 observações, se a potência de teste é 0.25, e 651 observações se a potência é elevada para 0.95. Para d=0.5 *n* varia entre 14 e 105 observações e para d=0.8 n varia entre 6 e 42 observações. Assim, *aumentar a potência do teste implica aumentar o número de observações em cada amostra.*

Finalmente, analisemos a relação entre a dimensão das amostras e o nível de significância α. Observamos na Tabela 2 que para os mesmos níveis de potência de teste e ME a dimensão das amostras deverá ser maior no caso em que α=0.01. Assim, para *diminuir o nível de significância há que aumentar a dimensão das amostras*, pois *para manter o mesmo nível de* n, *baixar alfa implica baixar a potência do teste*, como se verifica no nosso exemplo: se α=0.05, d=0.8 e 1-α=0.5, então n=13 enquanto que, para

manter n=13 com α=0.01 e d=0.8, 1-α baixa para metade (0.25).
Na realidade, querendo manter uma determinada potência de teste, o controlo directo do investigador recai, essencialmente, sobre a dimensão da sua amostra, visto que, regra geral, não é do seu interesse trabalhar com alfas muito elevados. Um controlo mais indirecto advém de todas as preocupações com possíveis fontes de variação estranhas ao seu estudo (redução de s^2).

Tipos de análise da potência de teste.

Existem fundamentalmente dois caminhos a seguir: 1) o de determinar a potência do teste após a realização da experiência, procedimento que se designa por análise de potência *post hoc* ou a *posteriori* ou 2) o de planear a experiência para assegurar uma potência de teste especificada *a priori*. Estes dois tipos de análise, em última instância, traduzem ora uma procura de mero conhecimento da potência associada ao teste realizado, ora uma procura de exercer controlo sobre a potência do teste a realizar.

Análise «a posteriori».

Quando, ao realizar um teste estatístico, decidimos não rejeitar Ho a um determinado nível de α, a tendência é para inferir a ausência do fenómeno em estudo. O cálculo da potência do teste realizado dar-nos-á indicações sobre a validade dessa inferência. Esta análise de potência é realizada *a posteriori*, ou seja, após se ter realizado o estudo e levado a cabo a análise estatística dos seus resultados. O investigador realiza-a para conhecer a probabilidade de detectar um efeito, caso ele realmente exista. A não-rejeição de Ho, podendo indicar ausência de efeito, pode estar associada a uma probabilidade suficientemente reduzida para detectar esse efeito, ainda que este estivesse presente na população estudada.
Consideremos o caso em que recorremos às tabelas de Cohen. Pré-estabelecemos o nível de α e a dimensão das amostras, n (que foi definida pelo número de observações a realizar durante a experiência). A *ME* é calculada usando a expressão apropriada ao teste estatístico, que é realizado e pode ser quantificada por um dos índices anteriormente referidos. Embora esta estimativa da *ME* deva ser considerada caso se pretenda estudar a presença ou ausência de um efeito, existem outras opções. Assim, por exemplo, se a *ME* estimada for pequena e ao investigador apenas interessar a detecção de efeitos de elevada magnitude, será do seu interesse demonstrar que o teste tem a potência suficiente à detecção dessa magnitude. Neste caso, estabelecerá a *ME* por um outro processo que não o de estimação (ver Análise *a priori*).
Conhecidos α, n e *ME* a potência do teste estatístico realizado pode ser determinada por consulta mais ou menos directa das tabelas de Cohen. Esta leitura nem sempre é directa, chegando mesmo a necessitar, por vezes, de outro tipo de informação sobre a análise em curso. É o caso do estudo onde não se realizaram igual número de observações por célula/amostra/grupo, sendo necessário determinar o valor de

n com que se vai consultar as tabelas (ver em Cohen 1988, as respectivas soluções). Por exemplo, para uma análise de variância (modelo de efeitos fixos onde os dados não violam os pressuposto de normalidade e homocesdasticidade), a determinação da potência necessita de um outro parâmetro de entrada nas tabelas de Cohen e que está relacionado com o número de níveis ou grupos de cada factor. O valor desse parâmetro corresponde ao número de graus de liberdade associados a cada factor.

Análise «a priori».

Ao fazer o planeamento de um estudo empírico, com vista a estudar um dado fenómeno (efeito), é desejável e apropriado controlar a potência que terá o teste que irá ser realizado para a análise dos dados. Ao proceder à análise da potência de teste, associada a um dado plano experimental, pode-se concluir, por exemplo, que a potência é de tal modo reduzida que a experiência terá que ser conduzida com um número de observações bem maior, ou até, que o esforço necessário não justifica prosseguir com esse plano experimental. No entanto, não ter em conta a potência de teste pode levar a que, na planificação do estudo, se cometa um erro relativamente ao número de sujeitos a estudar, erro esse que pode ser irrevogável e pôr em causa todo o trabalho realizado.

Na análise de potência *a priori*, o investigador especifica quais os riscos de ocorrência dos erros Tipo II e Tipo I que pretende admitir, bem como qual a magnitude do efeito (ME) que está a estudar. Estabelecidos, *a priori*, os níveis de α, da *ME* (padronizada) e da potência de teste, o investigador vai determinar *n*, a única variável nesta situação, utilizando por exemplo as tabelas construídas por Cohen (os gráficos de Pearson e Hartley, 1951, são outra via possível). Note-se que no caso deste *n* pré-estabelecido não vir a ser concretizado (por dificuldades de ordem prática) deve proceder-se a uma reanálise da potência (*a posteriori*).

Em termos práticos sabemos que, regra geral, o efeito a estudar é de magnitude desconhecida, e que é difícil obter a dimensão das amostras requerida para assegurar uma determinada potência de teste. Levanta-se, assim, de imediato a questão «Que fazer nestas situações?» Determinar diferentes valores de *n*, variando os níveis de ME, α e 1-ß, pode permitir ao investigador encontrar uma solução de compromisso que procure compatibilizar a dimensão da amostra necessária a níveis de α e potência de teste que considere aceitáveis. Esta solução de compromisso envolve, igualmente, uma análise cuidada dos objectivos do estudo, tendo por referência alguns pontos que passamos a discutir.

A Potência de Teste sob Controlo do Investigador.

Cabe ao investigador, que pretende exercer um controlo sob a potência das análises estatísticas a levar a cabo com os seus dados, analisar cuidadosamente as características do seu estudo, nomeadamente no que diz respeito aos

parâmetros da função potência de teste. Consideremos em primeiro lugar o parâmetro *ME*.

A estimação do parâmetro magnitude de efeito, que estamos a estudar, é, regra geral, apontada como o maior obstáculo a uma análise *a priori*, da potência dos testes. Uma estimativa poderá ser obtida i) com base em estudos anteriores realizados para as mesmas variáveis ou ii) com base em resultados de estudos piloto que se conduzam para o efeito. Na completa ausência destes guias o investigador pode ainda iii) apoiar-se na literatura disponível que, de alguma forma, sugira níveis de grandeza para o índice em questão (Cooper e Findley, 1982, sugerem no campo da Psicologia efeitos de magnitude média) ou, em última instância, iv) atribuir um nível de magnitude que considere verosímil.

Daqui se depreende uma distinção entre os estudos orientados por uma hipótese específica, derivada directamente de uma teoria, e os estudos com um carácter essencialmente exploratório. O investigador que se enquadra numa corrente teórica específica, com uma tradição de estudos associada, deve examinar atentamente a literatura, de forma a se aperceber da magnitude dos efeitos frequentemente detectados e, indirecta ou directamente, relacionados com o seu. Para o efeito, qualquer metanálise já realizada no campo será de grande utilidade. Estudos exploratórios, por outro lado, podem ter de se basear em «palpites» menos fundamentados. No entanto, recomendamos igualmente uma análise atenta à literatura que referencie variáveis da mesma natureza daquelas que o investigador vai analisar no seu estudo.

Uma outra distinção de tipos de estudo é relevante para estas considerações: a dos estudos com consequências interventivas. Estes estudos devem ser planeados relativamente a magnitudes que diferenciem a decisão associada a diferentes tipos de acção. Tomemos, como exemplo, um estudo de levantamento de necessidade de implementação de uma acção de formação numa empresa. O gestor decidirá pelo envolvimento numa acção formativa, apenas se um dado conjuntos de indicadores apontarem a existência de uma «grande necessidade». A diferença entre o que o gestor concebe ser «grande necessidade» e o que considera não ser «grande necessidade», é que deve ser considerada como índice de magnitude do efeito estudado, no cálculo da dimensão da amostra a utilizar no estudo. Note-se que, deste modo, a não rejeição de Ho apenas refere que a diferença não alcança a magnitude requerida (podendo existir diferença ou até não existir).

Caso o investigador não possa fazer uma estimativa do que será a magnitude do efeito a estudar poderá, pelo menos, definir se esta será pequena, média ou grande, atendendo assim às escalas funcionais de Cohen (1988). Por analogia, podemos comparar a potência de teste, relativamente à magnitude de efeito, com a potência de uma lente relativamente ao objecto a observar: quanto menor o objecto/efeito, maior a potência necessária para detectar a sua presença. Assim, enquanto ao pesquisador de bactérias oferecemos um microscópio e ao detector de pulgas uma lupa lembramos que, não faz sentido algum a quem procura um

elefante fornecer qualquer destes instrumentos. Mas, ao mesmo tempo, há que ter em conta que um teste de fraca potência, embora sendo capaz de detectar efeitos de grande magnitude, pode confundir a presença de uma bactéria com a sua ausência.

Esta necessidade de estimar a magnitude dos efeitos a estudar, dá grande relevância quer aos estudos metanalíticos, quer aos estudos exploratórios (que visam servir de suporte a um estudo de teste de hipóteses específicas). Deste modo, o leitor aperceber-se-á do serviço que pode prestar à comunidade científica se apresentar nos seus relatórios a magnitude dos efeitos estudados.

O parâmetro α não necessita de considerações sobre estimação, visto ser determinado previamente pelo investigador. Lembramos, no entanto, que este deve ter em consideração as implicações práticas de um erro Tipo I e de um erro Tipo II nas conclusões do seu estudo. Assim, consideremos o caso em que num determinado estudo o investigador vai formular uma Ho que «espera» vir a rejeitar (a hipótese de estudo coincide com H1). Nesta situação pode ser apropriado tomar uma atitude conservadora relativamente a α e, como tal, atribuir baixos níveis de significância, como os recomendados α=0.05 ou α=0.01. Se a decisão vem a ser rejeitar Ho então, condicional a Ho ser verdadeira, *apenas pode estar a cometer-se um erro Tipo I*, cuja probabilidade é conhecida a partir do momento em que se especificou α. Caso contrário, é aquele em que se «espera» não rejeitar Ho (a hipótese do estudo coincide com Ho). Nesta situação é óbvio que se deve aumentar a sensibilidade do teste, estabelecendo maiores níveis de potência de teste e, eventualmente, relaxando o controlo de alfa. Assim, novamente a distinção entre tipos de estudos condiciona a decisão sobre α. Nos estudos guiados teoricamente, a decisão de arriscar mais ou menos um dos tipos de erro, deriva directamente da hipótese em estudo (caso esta coincida com Ho ou com H1), e nos estudos exploratórios, ou outro tipo de estudos, onde os efeitos esperados sejam igualmente vagos, deve-se assumir um compromisso entre os dois tipos de erro: Cohen (1992), sugere a selecção do nível de potência de teste correspondente a níveis de alfa da ordem dos 0.10.

Nos estudos com implicações práticas, devem analisar-se as consequências associadas às medidas/acções que derivem das conclusões do estudo. Um estudo que pretenda discernir a existência de défices de aprendizagem num grupo de crianças com o fim de decidir por uma intervenção suplementar, deve preocupar-se mais com erros de Tipo II do que de Tipo I, porque concluir, erradamente, que existem défices (erro de Tipo I) pode representar uma acção interventiva junto de quem passava bem sem ela, enquanto que concluir erradamente a não existência de défices poderia conduzir à decisão de não intervenção, o que teria como consequência a não resolução de um problema com graves consequências no futuro. Pelo contrário, num estudo ergonómico que concluísse a não existência de uma associação entre a qualidade das condições de trabalho e o número de idas a uma consulta médica

desses trabalhadores, a preocupação deveria ser sobretudo com os erros de Tipo I. Isto porque, apenas face à conclusão de presença do efeito, se associavam medidas urgentes.

O parâmetro *n* é de todos os parâmetros referidos aquele que é mais maleável e, portanto, aquele sobre o qual o investigador vai exercer um maior controlo. A sua maleabilidade advém do facto de ser estabelecido pelo investigador (o que não acontece com *ME*) e de não ter influência directa sobre outro tipo de erro (o que não acontece com α). A regra, como já foi referido, é de que: quanto maior o número de observações realizadas, maior a potência dos testes que lhes estão associados. No entanto, os custos em tempo, recursos, etc., nem sempre são justificáveis. Note-se, porém, que em estudos que envolvem elevados recursos económicos é necessário assegurar uma elevada probabilidade de identificar um fenómeno que seja real donde, um *n* elevado pode ser economicamente justificado.

A preocupação com a dimensão da amostra, que aqui referimos, passa ao lado da questão de representatividade dessa amostra. A representatividade requerida pela procura de validade ecológica (i.e., capacidade de generalização dos resultados à população em estudo) relaciona-se, entre outros cuidados, com o recurso a uma amostra de dimensão elevada, associada a um processo de amostragem adequado (com base na Teoria de Amostragem que nos oferece algoritmos para o cálculo dessa dimensão - ver, por exemplo, Rea & Parker, 1992). Pelo que, estudos com objectivos essencialmente descritivos têm, assim, um duplo intuito de validade, ao aumentarem a dimensão das amostras com que trabalham. Por sua vez, os estudos experimentais (laboratório), procurando detectar a presença ou ausência de um efeito postulado teoricamente, preocupam-se com a capacidade de generalização das suas conclusões e, não tanto, dos seus resultados. Assim sendo, a dimensão das amostras com que se trabalha está meramente dependente da capacidade de, com essas amostras, se detectar um efeito quando este está presente na população subjacente, ou seja, está directamente relacionada com a potência do teste realizado para a hipótese formulada.

A questão, em torno do estabelecimento de ß, está em saber como se pode interpretar «maiores níveis de potência» ou «níveis de potência razoáveis». Realmente não existe acordo entre os investigadores sobre esta questão, como existe, por exemplo, na atribuição dos níveis de alfa ($\alpha = 0.05$ como valor mais usado). Cohen (1988) estabelece como convenção a necessidade de se reduzir a probabilidade de erros Tipo II a 0.20 e, portanto, realizar testes cuja potência seja à volta de 0.80. Keppel (1991) considera 0.80 não só como um nível aceitável, mas também realista. Kraemer e Thiemann (1987) sugerem uma oscilação de 0.1 em torno desse valor, estabelecendo a necessidade de a potência de um teste se encontrar no intervalo [0.7; 0.9]. O aumento da potência de teste, acima destes valores, implicaria o recurso a amostras de dimensão muito elevada e, abaixo deste valor, torna difícil qualquer interpretação de um

resultado não significativo. Como o autor refere (Cohen, 1992) tomando em conjunto $\alpha = 0.05$ e a potência de teste de 0.80, a razão b:a é de 4:1 (0.20 para 0.05), pelo que arriscamos quatro vezes mais um erro de Tipo II do que um erro de Tipo I.

Conclusão.

Neste artigo pretendemos salientar os seguintes aspectos:

1) Ignorar os erros de Tipo II é particularmente perigoso quando a análise estatística, ou teste estatístico, leva o investigador a não rejeitar Ho. A não rejeição da hipótese nula pode resultar não do facto de esta hipótese ser realmente verdadeira, mas do facto de as observações serem insuficientes para declarar inaceitável Ho. Assim, antes de podermos afirmar validamente que uma hipótese nula é verdadeira, devemos sempre assegurarmo-nos de que o risco de segunda espécie, ß, também é suficientemente pequeno.

2) Um $\alpha=0.05$ ou 0.01 nem sempre é justificável. Os valores de alfa em torno de 5% ou de 1% são quase valores míticos que regem a validade dos nossos estudos e, embora tenham algum sentido quando queremos fazer face a problemas de Tipo I, devem ser ponderados quando a preocupação tem, igualmente, de recair sobre os erros de Tipo II. Os valores 1% e 5% são valores típicos ou padrão, ou seja, valores com os quais devemos confrontar as nossas decisões. No entando, não nos devemos esquecer que *God loves the 0.06 nearly as much as the 0.05* (Rosnow & Rosenthal, 1989, p. 1277) e que quando procuramos aceitar Ho, Deus até que deve amar mais os valores superiores a 0.05.

3) Existe uma dupla vantagem em se trabalhar com amostras de dimensão elevada: representatividade e potência de teste. No entanto, a escolha da dimensão da amostra deve ser cuidada, não esquecendo que existem situações em que trabalhar com grandes amostras se demonstra ser totalmente desapropriado.

4) Não deve ser confundida (como muitas análises o fazem) significância estatística (que tem a ver com probabilidade de se verificar o efeito por acaso) e significância prática ou psicológica (que tem a ver com a magnitude do efeito estudado). Grandes amostras conduzem-nos a testes muito potentes, capazes de detectar pequenas magnitudes de efeito: o efeito é estatisticamente significativo. Se o é ou não, em termos práticos, cabe ao investigador decidir: um efeito isolado em laboratório pode ser previsto pela teoria até em magnitude reduzida; no entanto, mudanças de política em educação, envolvendo elevados custos, podem ser apenas justificadas por efeitos de magnitude elevada.

5) Existe consenso sobre a importância de analisar e apresentar os resultados da potência de teste (Schafer, 1993), e uma vasta literatura de apoio para modelos paramétricos e não-paramétricos, da qual destacamos os livros de Cohen (1988) e de Kraemer & Thiemann (1987). Referências mais específicas podem ser igualmente encontradas, por exemplo: Koele (1982), que apresenta soluções para o cálculo da potência de teste para modelos ANOVA de efeitos aleatórios e mistos

(ver também Guenther, 1964 para efeitos aleatórios); e Barcikowski (1973), que apresenta tabelas para determinação da dimensão das amostras no caso dos modelos de efeitos aleatórios. No entanto, mesmo quando não se encontram publicadas as expressões analíticas para o cálculo da potência de um qualquer teste, é sempre possível recorrer a simulações que vão permitir gerar as respectivas tabelas de potência (Stephens, 1974).

Por último deixamos duas sugestões aos leitores mais curiosos no assunto:

i) No caso de simples ensaios de significância de parâmetros (Fisher, 1932), em vez de se recorrer à formulação da hipótese Ho: parâmetro=0 (o *default* deste tipo de ensaio), será adequada a formulação de Ho como *p*=magnitude do parâmetro que determina a decisão de lhe atribuir significado prático ou não. No caso do teste de hipótese simples (em que o espaço do parâmetro tem apenas dois elementos: ver Murteira, 1990), para detectar apenas efeitos de magnitude pretendida, será adequada a formulação do teste (de Neyman-Pearson, 1928) que especifica em Ho a magnitude exata do efeito que este pretende detectar;

ii) Para fazer face ao problema da dimensão de *n*, ao estabelecer antecipadamente α e ß, poderá, em alguns casos (nomeadamente o de comparação de duas médias), encontrar solução no recurso a testes sequenciais (Wald, 1947; McPherton & Armitage, 1971).

Mesmo tendo este artigo algum efeito sobre a atitude do leitor relativamente ao tratamento dos seus dados, tudo parece indicar que não afectará o seu comportamento. Na realidade, alguns artigos (como: Cohen (1992) e Sedlmeier & Gigerenzer (1989)) focam a questão da ineficácia dos alertas constantes da literatura para a preocupação com a magnitude da probabilidade de erro Tipo II no aumento da potência de teste das análises dos resultados das investigações. Isto é, apesar de a maioria dos metodólogos e investigadores concordar com a importância e necessidade de controlo da probabilidade de erro Tipo II, este tipo de controlo não é exercido quando se analisam os resultados de investigações publicados. Talvez Bem & Honorton (1993) tenham razão quando (a título de piada) levantam a hipótese de o comportamento do investigador ser «ultimately determinated by whether one was more severely punished in childhood for Type I or Type II errors» (p. 11).

Esperamos, porém, que o presente artigo contribua, de algum modo, para uma referência quase que universal ao livro de Cohen (1988); para que os investigadores se abstenham de interpretar resultados nulos (testes onde não se rejeita a hipótese nula), caso desconheçam a potência dos testes levados a cabo e, ainda, para se eliminarem as confusões de interpretação entre a significância estatística e a magnitude do efeito estudado.

[3] Note-se que as normas de publicação em revistas da APA (4.ª Ed.) referem a necessidade de se reportarem os valores de potência dos testes levados a cabo.

Referências

Barcikowski, R. S. (1973). Optimum sample size and number of levels in a one-way random-effects analysis of variance. *Journal of Experimental Education*, 41, 10-16.

Bem , D.J., & Honorton, C. (1994). Does psi exist? Replicable evidence for an anomalous process of information transfer. *Psychological Bulletin*, 115 (1), 4-18.

Campbell, D.T. (1957). Factors relevant to the validity of experiments in social settings. *Psychological Bulletin*, 54, 297-312.

Cook, T. D. & Campbell, D.T. (1979). Quasi-experimentation. Design & analysis issues for field settings. Boston: Houghton Mifflin Company.

Cohen, J. (1962). The statistical power of abnormal-social psychological research: a review. *Journal of Abnormal Social Psychology*, 65, 145-153.

— (1988). *Statistical power analysis for the behavioral sciences*, (2nd ed.). Hillsdale, N.J.:Lawrence Erlbaum

— (1992). Statistical power analysis. *Psychological Science, 1* (3), 98-101.

Cooper, H. & Findley, M. (1982). Expected effect sizes: estimates for statistical power analysis in social psychology. *Personality and Social Psychology Bulletin, 8* (1), 168-73.

Fisher, R.A. (1932). *Statistical methods for research workers*. Edinburg: Oliver & Boyd.

Garcia-Marques, T., & Azevedo, M. (1994). A inferência estatística múltipla e o problema de inflação do alfa. *Psicologia* (para publicação).

Goldstein, R. (1989). Power and sample size via MS/PC-DOS computers. *The American Statistician*, 43, 253-60.

Guenther, W. C. (1964). *Analysis of variance*, Englewood Cliffs, NJ: PrenticeHall.

Keppel, G. (1991). *Design and analysis: A researcher's handbook* (3rd ed.). Englewood Cliffs, NJ: Prentice-Hall.

Koele, P. (1982). Calculating power in analysis of variance. *Psychological Bulletin*, 92, 513-16.

Lawrence, B. M. (1990). *Understanding significance testing*. Beverly Hills,CA: SAGE

McPherson, G.K., & Armitage, P. (1971). Repeated significance testes on accumulating data when the null hypothesis is not true. *Journal of the Royal Statistical Society, A*, 134, 15-25.

Murteira, B. J. F. (1990). *Probabilidades e estatística* (vol. II). Lisboa: McGraw-Hill.

Neyman, J. & Pearson, E. S. (1928). On the use and interpretation of certain test criteria for purposes of statistical inference. *Biometrika*, 20A, 175-240, 263-94.

Pearson, E. S., & Hartley, H. O. (1951). Charts for the power function for analysis of variance tests, derived from the non-central F-distribution. *Biometrika*, 38, 112-30.

Peterman, R. M. (1990). Statistical power analysis can improve fisheries research and management. *Canadian Journal of Fishing and Aquatic Science*, 47, 2-15.

Rea, L. M., & Parker, R. A. (1992). *Designing and conducting survey research. A comprehensive guide*. San Francisco: Jossey-Bass.

Rosnow, R. L., & Rosenthal, R. (1989). *American Psychologist*, 44 (10), 1276-84.

Schafer, W. D. (1993). Interpreting statistical significance and nonsignificance. *The Journal of Experimental Education*, 61(4), 383-87.

Scheffé, H. (1959). *Analysis of variance*. New York: Wiley.

Sedlmeier, P., & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of the studies? *Psychological Bulletin*, 105, 309-16.

Stephens, M. A. (1974). EDF statistics for goodness of fit and some comparisons. *Journal of the American Statistician Association*, 69, 730-37.

Wald, A. (1947). *Sequential analysis*. New York: Wiley.

Abstract. - Statistical analysis in research allow us to decide about the presence or absence of an effect in our data. How confident are we about this decision? The power of the statistical tests, the relation between Type I and Type II errors with the acceptance of Ho, statistical significance, pragmatic significance and representativity are covered in this article.