ARTIGO TEÓRICO

Interpretação crítica dos resultados estatísticos: para lá da significância estatística

Critical interpretation of statistical results: beyond statistical significance

Interpretación crítica de resultados estadísticos: mas allá de la significancia estatística

Luís Manuel de Jesus Loureiro*
Manuel Gonçalves Henriques Gameiro**

Resumo

Este artigo tem como objectivos a descrição e análise do processo de tomada de decisão (interpretativo e conclusivo) em estatística, quando se recorre a testes de hipóteses. É apresentada a lógica do processo decisório e os conceitos de erros tipo I e II, poder do teste e magnitude do efeito (*effect size*), demonstrados a partir da medida «d» de magnitude do efeito.

O artigo dirige-se a duas situações críticas: resultados com significado estatístico mas sem significado clínico efectivo; e resultados sem significado estatístico mas que se devem a problemas relacionados com a natureza dos fenómenos, *effect size* pequeno e reduzido tamanho da amostra.

Palavras-chave: erros de inferência; tamanho do efeito; poder do teste; significância.

Abstract

This paper aims to describe and analyse the decision-making process (conclusive and interpretative) in statistics when performing hypothesis tests. The logic of the decision process is presented, as well as the concepts of type I and type II errors, power of the test and effect size. We also make demonstrations of effect size «d» measure.

This paper targets two critical situations: results of statistical significance but no real clinical significance, and results without statistical significance but which are due to problems related to the nature of the phenomena, small effect size and small sample size.

Keywords: inference errors; effect size; power test; significance.

Resumen

Este artículo tiene el doble objetivo de describir y el analizar el proceso de toma de decisión (interpretativo y conclusivo) cuando se recurre en estadística a tests de hipótesis. Se presenta a continuación la lógica del proceso decisorio y los conceptos de errores de tipo I y II, la potencia de la prueba y el tamaño del efecto (*effect size*), demostrados a partir de la medida «d» de tamaño del efecto.

El artículo se orienta hacia dos situaciones críticas: resultados con significado estadístico aunque sin significado clínico efectivo; y resultados sin significado estadístico pero que se deben a problemas relacionados con la naturaleza de los fenómenos, *tamaño del efecto* pequeño y reducido tamaño de la muestra.

Palabras clave: errores de la inferencia, tamaño del efecto, prueba de potencia, significancia.

* Professor Adjunto da Escola Superior de Enfermagem de Coimbra (ESEnfC – UCPSMP). Doutorado em Saúde Mental [luisloureiro@esenfc.pt].
** Professor Coordenador da ESENFC – UCPSMP. Mestre em Ciências da Enfermagem [mgameiro@esenfc.pt].

Recebido para publicação em: 21.08.10
Aceite para publicação em: 04.02.11

Introdução

É reconhecido o impacto que tiveram, nos domínios académico e científico, dois artigos publicados na década de noventa do século precedente, um da autoria de Cohen (1994), com o título «*The earth is round (p<.05)*», o outro de Johnson (1999) sob o título «*The insignificance of statistical significance testing*». Eram apontadas, como críticas e insuficiências, quer a ausência do cálculo de diferentes medidas (ex: Intervalos de confiança) que se traduzia em interpretações abusivas e conclusões inocentes dos resultados dos testes de significância estatística (Wilkinson e APA, 1999), quer ainda uma subserviência desmedida ao *p-level* (ou *sig.*) dos *softwares* estatísticos, tomadas como sinónimos e garante do valor científico dos resultados obtidos, ou mesmo, da qualidade das hipóteses formuladas e desenho utilizados (Glasser, 1999).

Bastará ainda hoje folhear as publicações científicas ou assistir resignadamente às comunicações em congressos e simpósios, e todos os outros meios de difusão científicos, para ouvir quer as menções sonantes às relações ou diferenças estatisticamente significativas encontradas, escoltadas pela célebre simbologia $p<.05$, $p<.01$ e $p<.001$ e respectivos asteriscos ou estrelinhas [*; **; ***], quer a referência «*ns*», ou $p>.05$, sinónimo de um resultado não significativo de um ponto de vista estatístico.

Em ambos os casos, encontrando-se ou não diferenças/relações estatisticamente significativas nas amostras observadas, não basta acenar com os *p-level* (ou *sig.*), considerando satisfatório quando os resultados vêm ao encontro do que é expectável, ou então, diminuindo/desvalorizando-os (resultados) quando o não são, relevando o seu significado efectivo. No Quadro 1, podem observar-se aquelas que são as reacções dos investigadores, sendo que, na generalidade das vezes, apenas se dá importância às significâncias estatísticas, omitindo a significância prática.

QUADRO 1 – Reacções do investigador às significâncias estatística e prática

Importância prática das diferenças observadas	Significância Estatística	
	Não significativa	Significativa
Não é importante	Alegria	Aborrecimento
É importante	Tristeza	Euforia

Fonte: Nester (1996) *cit. in.* Johnson, (1999, p. 776)

Em ambos os casos, podemos estar perante aquilo que Johnson (1999) designou como resultados estatisticamente significativos e substancialmente insignificantes ou, então, quando não são estatisticamente significativos, o investigador «escava» as insuficiências e erros, pensando muitas vezes que agindo dessa maneira lhes consegue atribuir significado (Johnson, 1999). É comum, neste último caso, o investigador "apontar baterias" ao tamanho da amostra, ou questionar da sua aleatoriedade e representatividade. Relativamente às significâncias, estatística e prática, considerando o tamanho da amostra (n), são apresentadas no Quadro 2 as considerações comuns dos investigadores.

QUADRO 2 – Considerações acerca do tamanho da amostra (*n*) no contexto das significâncias (estatística e prática)

Importância prática das diferenças observadas	Significância estatística	
	Não significativa	Significativa
Não é importante	*n* é bom	*n* é muito grande
É importante	*n* é muito pequeno	*n* é bom

Fonte: Nester (1996) *cit. in.* Johnson, (1999, p. 776)

Os dois quadros (1 e 2) reflectem uma atitude por parte dos investigadores que tendem a ver o valor e qualidade dos resultados a partir das significâncias estatísticas (Glasser, 1999), descurando o conceito de significância prática. Assim, e na afirmação de que uma relação ou diferença observada é estatisticamente

significativa ou que os resultados vêm ao encontro ou não, dos de *a*, *b* ou *c*, reflecte-se normalmente uma insuficiência, quer quanto ao alcance dos resultados, quer quanto aos limites e fragilidades do próprio estudo, o que se pode traduzir em leituras aligeiradas e acríticas.

Então, para lá do esforço entusiasmado dos investigadores, verificado no relevo que é dado ao debitar de resultados considerados estatisticamente significativos, o que subsiste destas declarações mais ou menos convictas? Qual o fundamento destas decisões e a sua relevância no contexto da investigação? Qual o alcance e significado substantivo dos resultados tidos como sendo, ou não, estatisticamente significativos?

A resposta a algumas destas questões passa por entender quer a lógica do processo de tomada de decisão estatística (erros comuns envolvidos tipos I e II), quer, inclusive, o manuseamento de outras medidas estatísticas, como a de tamanho ou magnitude do efeito (*effect size*) e as análises do poder dos testes (relacionadas com o erro tipo II). Torna-se, pois, fundamental reconhecer que significância estatística não é sinónimo de significância prática ou clínica (Kendall, 1999; Tallmadge, 1977), isto porque, um resultado pode ser estatisticamente significativo e não ter relevância, sendo de reter que a substancialidade não se esgota nos valores de *p* obtidos. Os objectivos deste artigo passam, não só, por entender o processo decisório em estatística, mas, sobretudo, contribuir para uma interpretação crítica dos resultados estatísticos obtidos, não se cingindo apenas aos valores das significâncias estatísticas.

O processo de decisão em estatística

Quem recorre à estatística como ferramenta de tomada de decisão depara-se, antes mesmo de calcular qualquer medida ou teste estatístico ou ainda no processo de ensino aprendizagem da disciplina de investigação, com o conceito de *erro*. Este é um conceito comum no campo da investigação científica, especialmente quando se recorre aos métodos quantitativos, onde a estatística se assume como uma ferramenta fundamental. Por exemplo, quando se utiliza a média amostral (estatística) como estimador da média populacional (parâmetro), fala-se de *erro padrão da média* ($\hat{s}/\sqrt{n}$)[1] com o objectivo de caracterizar a precisão desse estimador. De facto, os *erros aleatórios* querem significar aqueles erros que derivam do acaso, como é o caso do *erro amostral*. A questão não está tanto em trabalhar com estes erros, mas na forma de os reduzir ou minimizar, admitindo que nunca se poderá decidir sem um erro associado, no entanto, ao erro amostral acresce ainda o erro associado à decisão.

Na investigação, seja no âmbito da Enfermagem ou noutros campos do saber, trabalha-se tradicionalmente com testes de hipóteses. Partindo de um corpo teórico formula-se uma ou várias hipóteses e depois, com base nos dados empíricos, tomam-se decisões. Face à hipótese o que poderá o investigador decidir? Com que base lógica toma uma decisão?

Um exemplo simples poderá servir para compreender a formulação lógica das hipóteses e a sua ligação com o raciocínio estatístico. Tome-se a proposição de que *todos os cisnes são brancos*. Pode-se demonstrar sem erro (provar) esta afirmação? Naturalmente que não, pois não se conhecem todos os cisnes e pode-se colocar a hipótese de que existam alguns de outra cor. A única certeza existente prende-se com o facto de todos os cisnes que se observam serem brancos, ou seja, a certeza que advém da experiência. Mas se aparecesse um cisne azul, isso bastaria para provar que se estava errado, ou seja, para falsificabilizar ou provar como errada a proposição.

Assim, a regra da lógica subjacente é: nenhuma afirmação ou declaração pode ser provada por inferência como verdadeira, mas apenas como falsa. Faz-se recurso a um procedimento lógico designado por *modus tollens,* também denominado como procedimento de falsificação. Este procedimento assenta no facto de uma observação singular poder conduzir à conclusão de que a premissa é incorrecta. A utilização do *modus tollens* na estatística exige a existência de duas hipóteses, a hipótese nula ou estatística (H_0) e a hipótese alternativa ou contra hipótese (H_1).

Será utilizado o teste *t* de comparação de médias em grupos independentes como exemplo para explicitar estas ideias, sendo também que este teste é um dos testes mais utilizados na investigação científica publicada.

---

[1] $\hat{s} = s\sqrt{\frac{n}{n-1}}$: melhor estimativa do desvio padrão populacional (σ) a partir do desvio padrão amostral (s).

As hipóteses de nulidade e a hipótese conceptual ou de investigação

A hipótese nula (H_0) é a afirmação *a priori*, aquela que vai ser posta à prova, isto é, utilizam-se os resultados para decidir no sentido de a rejeitar (assumindo que é falsa) ou não rejeitar (assumindo que é verdadeira). É traduzida simbolicamente por $\mu_1=\mu_2$. Este processo é análogo à presunção de inocência do direito penal. Até prova em contrário, o réu é inocente; face às provas, o juiz ou jurados decidem: culpado ou não culpado. A inocência é traduzida em estatística pela condição de igualdade ($\mu_1=\mu_2$). É sobre esta relação entre parâmetros que se toma uma decisão, ou se rejeita, ou não se rejeita a igualdade das médias e, quando se rejeita, opta-se pela hipótese de investigação ($\mu_1\neq\mu_2$) como uma justificação *mais adequada* para explicar o fenómeno.

A hipótese é definida por Kerlinger (1980, p. 38) como «um enunciado conjectural das relações entre duas ou mais variáveis. Hipóteses são sentenças declarativas e relacionam, de alguma forma, variáveis a variáveis. São enunciados de relações, e (...) devem implicar a testagem das relações enunciadas». Ainda que a definição não refira que se podem formular hipóteses univariadas, repare-se que o conceito inclui o de *relação*, querendo significar ou definir o *comportamento* relativo das variáveis. Uma distinção importante, relativamente à utilização do conceito, é a de que quando falamos de hipótese estatística referimo-nos a uma relação matemática entre dois ou mais parâmetros populacionais, quando falamos de hipótese conceptual, entendida aqui como hipótese operacional ou de investigação, fazemos uma afirmação que prevê uma relação entre variáveis, por exemplo uma dependente e uma independente (Runyon *et al.,*1996). Todavia, na maioria das vezes a designação das variáveis como independentes e dependentes é indevida, para não dizer abusiva, pois nos tipos de estudos realizados não existe qualquer tipo de controlo da variável independente, ainda assim, é um bom meio para distinguir o estatuto das variáveis na investigação.

Ainda relativamente ao conceito de relação, este encontra-se aplicado regularmente em todos os tipos de hipóteses, no entanto, calculadas as medidas dos grupos amostrais e os valores dos testes de significância, verifica-se que não é apresentada qualquer referência à medida de relação, especificamente nos testes de diferença de médias ou, por exemplo, no teste de diferença de proporções do qui-quadrado (tabelas de contingência). Contudo, quando os investigadores recorrem ao coeficiente de correlação de Pearson, simbolizado a estatística pela letra «r», sabem que esta medida indica o grau ou força e sentido da relação linear entre duas variáveis quantitativas (*x* e *y*) na amostra. Se o valor do coeficiente é ou não estatisticamente significativo, avalia-se *a posteriori* através do teste de significância da correlação.

O coeficiente de correlação (r) tem associado um outro coeficiente, designado por *coeficiente de determinação* (r^2) que indica a proporção (%) de variação em *y* que é explicada pela variação em *x* e *vice-versa*. Assim, se r=.40, logo r^2=0.16, se r=.70, r^2=0.49 e assim sucessivamente. Se, quando é utilizado o coeficiente de correlação linear de Pearson para testar uma hipótese, os investigadores referem, algumas vezes, o valor do coeficiente de determinação para atestar a relação encontrada entre as variáveis na amostra, noutros testes existe uma resistência ao cálculo deste tipo de medidas. Por exemplo, se o investigador recorre ao coeficiente (r) e respectivo teste de significância para testar a hipótese de que *«existe relação entre a idade dos estudantes do ensino superior e a permissividade na sexualidade»*, sendo a amostra de 960 estudantes, e encontra um valor de r=.101, com p<.01, deve ser rigoroso na interpretação deste resultado, isto porque, apesar de a correlação ser estatisticamente significativa, ela não se reveste de significado prático ou substancial (r^2=.01), dado que, nesta amostra tão numerosa, a variância explicada é de 1%.

Nos testes de diferenças de médias ou no qui-quadrado, entre outros, como já se referiu, ocorre uma situação análoga, mas mais gravosa, pois ocultam-se muitas vezes os valores das estatísticas resumo e mesmo da estatística do teste, acenando-se em exclusivo com o valor da *significância* (*p)* associado ao teste calculado. É de referir que existem medidas para avaliar a força das relações, ou das diferenças, noutra tipologia de testes, como é o caso do ómega (ω) no teste *t*, o eta (η) na Anova de um critério (quando o factor não comporta uma ordem intrínseca) e o ϕ (phi) ou o V de Cramer, no Qui-Quadrado (χ^2), consoante as tabelas sejam de 2x2 ou IxC. Estas medidas são úteis em todas as circunstâncias, mas revelam-se fundamentais para a

crítica dos resultados quando as diferenças/relações são estatisticamente significativas (Murphy, 1998).

Quando o investigador enuncia, como hipótese do estudo, que «*o burnout está relacionado com a natureza dos cuidados prestados pelos enfermeiros*», imaginando que se avalia e compara o *burnout* de enfermeiros de dois serviços distintos (internamento de oncologia e serviço de urgências gerais), ao utilizar o teste *t* de diferença de médias, ele deve calcular posteriormente a respectiva medida de associação ou então, indicar a percentagem de variação no *burnout* dos enfermeiros que é devida à natureza do serviço onde trabalham. De modo prático responde a duas questões: a) há evidência para afirmar que os níveis de *burnout* são diferentes? b) Se sim, qual a substancialidade de tal diferença, ou, se não, a que se pode dever tal resultado?

Não se deve esquecer que os testes de hipótese são um conjunto de técnicas de inferência estatística que avaliam a probabilidade de determinados desvios, diferenças, ou relações registadas nas distribuições observadas, serem devidos ao acaso amostral ou à existência real na população, tendo por base, obviamente, observações realizadas em amostras (subconjuntos que se pretendem representativos da população). É também necessário relembrar que as amostras deverão ser não só representativas, para que se possa generalizar os resultados, mas também aleatórias, permitindo a aplicação da teoria das probabilidades com a aplicação das técnicas de inferência (as técnicas de inferência exigem amostras aleatórias).

O erro na decisão estatística – analogias

Como se referiu, a decisão estatística está sujeita ao erro, como é o caso das decisões que derivam da aplicação dos testes de hipóteses. Os testes baseiam-se nas regras da inferência negativa, ou seja, parte-se do pressuposto de que a hipótese nula é verdadeira, à semelhança do conceito de presunção de inocência no direito penal. Até prova em contrário, todo o arguido deve ser considerado inocente. Em analogia com os testes de hipóteses, a hipótese nula é verdadeira até que uma evidência suficientemente forte indique que essa afirmação é incorrecta, com baixa probabilidade de erro. Exemplificando: a hipótese H_0 será o acusado ser inocente de ter cometido o crime x, a contra hipótese H_1 será o acusado ser culpado de ter cometido o crime x.

No sistema judicial, para que um indivíduo seja condenado é necessário que se apresentem provas nesse sentido e estas devem ir para além de uma *dúvida razoável*. Face às provas obtidas, o juiz toma uma decisão: considera o acusado culpado ou não culpado. Se existem semelhanças entre o raciocínio jurídico e os testes de hipóteses, existem também diferenças, nomeadamente, no conceito de prova. A rejeição da hipótese nula não constitui prova, em definitivo, de que a hipótese alternativa seja válida, isto é, trata-se apenas de uma evidência, muitas vezes provisória, de que a hipótese nula é provavelmente incorrecta. Mesmo que ela tenha sido declarada como incorrecta, há possibilidade dela ser verdadeira.

Com os testes de hipóteses faz-se do seguinte modo: formula-se uma H_0 e, face ao conjunto de dados amostrais, toma-se uma decisão: rejeita-se ou não a hipótese nula. Relembre-se o exemplo já apresentado de que o "*burnout está relacionado com a natureza dos cuidados prestados pelos enfermeiros*" Simbolicamente, tem-se H_1: $\mu_1 \neq \mu_2$. No entanto, partindo do princípio de que nada há que justifique, *a priori,* que as médias sejam diferentes, portanto a H_0, hipótese a testar, será $\mu_1 = \mu_2$, e é sobre esta que se toma a decisão. Quando se rejeita é porque existe uma *evidência suficientemente forte* para considerar que, efectivamente, o nível de *burnout* é diferente consoante a natureza dos cuidados prestados pelos enfermeiros (relacionados com o local de trabalho). Quando não se rejeita é porque as diferenças observadas nas médias se podem dever ao acaso amostral e por isso mesmo se considera $\mu_1 = \mu_2$. Pode-se encontrar diferenças estatisticamente significativas entre as médias dos grupos ($p < .05$), sendo que, o trabalho posterior passa por analisar o alcance dessa significância.

Como se sabe, muitas vezes cometem-se erros nos tribunais. Não é a primeira vez que alguém que foi condenado à prisão é posteriormente colocado em liberdade, por se ter verificado entretanto que a pessoa não era culpada. Uma decisão do tribunal acarreta dois erros possíveis, pode-se libertar um criminoso ou prender um inocente. Na decisão estatística passa-se o mesmo. O investigador pode rejeitar a hipótese nula quando ela é de facto verdadeira (erro tipo I), ou então não a rejeita quando ela é uma afirmação falsa (erro tipo II). É quando se rejeita a hipótese nula que

aparecem as célebres expressões: $p<.05$, $p<01$ ou ainda $p<.001$.

Assim, relativamente aos erros, o erro tipo I[2] designado por alfa (α), define-se como a probabilidade de rejeitar a hipótese nula quando ela é verdadeira, enquanto o erro tipo II, designado por beta (β), define-se como a probabilidade de não rejeitar a hipótese nula quando ela é falsa. No quadro 3 estão indicados os dois tipos de erro e enquadradas as decisões estatísticas por analogia com a decisão do tribunal. Como se pode observar, quatro resultados são possíveis numa decisão.

QUADRO 3 – Decisões e erros associados - analogia com sistema judicial

A interpretação da realidade ⇒ / Os factos (a realidade) ⇓	Rejeitamos a H_0 **Declarado culpado** [O *burnout* é diferente]	Não rejeitamos a H_0 **Declarado não culpado** [O *burnout* não é diferente]
(H_0 é verdadeira) - (na realidade inocente) [O *burnout* não é diferente]	Erro tipo I (α)	Decisão Correcta
(H_0 é falsa) - (realmente culpado) [O *burnout* é diferente]	Decisão Correcta	Erro tipo II (β)

Qual dos erros é mais importante controlar e evitar? Na realidade, ambos são importantes, mas tradicionalmente dá-se mais importância ao erro tipo I, um pouco como no tribunal. Na dúvida, decide-se a favor do réu, isto é, na dúvida, é preferível libertar um criminoso a prender um inocente.

Como se controlam esses erros na estatística? O erro tipo I é controlado à partida, através do nível de significância do teste. O investigador decide o risco que quer correr para rejeitar a H_0. Os níveis de significância utilizados convencionalmente nas ciências sociais e humanas, como valor critério para a rejeição da hipótese nula, são 5% (0,05) e 1% (0,01), mas pode-se encontrar também 10% (0,10).

O controlo do erro tipo II é mais complexo e está relacionado com o poder dos testes estatísticos. O risco do erro tipo II é afectado por variados factores: tamanho das amostras, desenhos de investigação, relacionamento previsto entre variáveis, bem como, o tipo de teste que se utiliza. Normalmente quando se estabelece um critério rígido para o α, aumenta-se a probabilidade de cometer o erro tipo II.

Erro tipo II e Poder do teste

Em seguida são apresentados os resultados de dois estudos (Quadro 4) com base na mesma hipótese, realizados em duas grandes instituições hospitalares (A e B) de Portugal Continental, admitindo que as amostras são equivalentes relativamente às características sociodemográficas dos enfermeiros. Os resultados da aplicação dos testes (assumidos os pressupostos de distribuição normal da variável na população e a homogeneidade das variâncias) e as estatísticas resumo são apresentados no mesmo quadro. O nível de significância utilizado (erro tipo I) para ambos os testes foi de $\alpha = .05$.

Como pode observar do Quadro 4, a diferença das médias do *burnout* dos enfermeiros é estatisticamente significativa na instituição A ($t_{(138)} = -4.643; p = .000$), enquanto na instituição B, a diferença não o é ($t_{(58)} = -1.903; p = .062$).

QUADRO 4 – Estatísticas resumo e teste *t* de *Student* de comparação de médias do *burnout* dos enfermeiros

Local:	n	$\overline{x}$	$\hat{s}$	***sem***	t	***p***
Urgência da instituição A	70	20.34	3.45	.41	-4.643	.000
Oncologia da instituição A	70	22.94	3.17	.38		
Urgência da instituição B	30	20.65	3.27	.60	-1.903	.062
Oncologia da instituição B	30	22.23	3.16	.58		

[2] Tecnicamente é designado por nível de significância e simbolizado por α. Indica a área numa distribuição teórica de probabilidades que corresponde à condição (valor crítico) de rejeição da hipótese nula.

Antes mesmo de procurar analisar os resultados em cada um dos estudos relativamente à sua substancialidade, nomeadamente, calculando outras medidas que auxiliem no trabalho interpretativo, deve atentar nos resultados da instituição B. De facto, ao afirmar que a diferença não é estatisticamente significativa para um nível de significância de .05, pode-se estar a cometer um erro tipo II (Quadro 3), isto é, afirmar que as médias são iguais quando de facto não são, o mesmo é dizer que não se rejeita a hipótese nula quando ela é falsa.

O erro tipo II (β) tem associado a si o conceito de poder do teste ($1-\beta$). Este erro tem sido menosprezado na prática corrente da investigação em ciências sociais e humanas, mas o seu controle é importante. A sua omissão ou desconsideração pode constituir uma grave lacuna no processo de investigação, especialmente nos casos em que se assume que as diferenças não são significativas e todo o processo cessa aí, como se os resultados não tivessem significado ou o desenho de investigação não precisasse de ser repensado.

O poder do teste pode definir-se como *a probabilidade de correctamente rejeitar uma falsa* H_0 ou, dito de outra forma, *a probabilidade de rejeitar a hipótese nula quando ela é falsa*. Assim, enquanto no erro tipo I se utiliza 5% ou 1%, é comum referir que o teste deve ter um poder (potência) de 80% (para β de 20%). No caso das ciências sociais e humanas, refere-se que o poder nunca deverá ser menor de 50% (Hill e Hill, 2000). Então, quando não se rejeita H_0 sem mais considerações, e se tem um teste com *fraco* poder, isso pode mostrar alguma ingenuidade por parte do investigador.

Face aos resultados dos dois estudos, quais as reações mais comuns nos investigadores? Naturalmente que as descritas por Johnson (1999), e referidas no Quadro 1, ainda que o investigador se cinja à leitura da significância estatística. O que distingue estes dois resultados de facto, para lá da significância e do tamanho da amostra? São esses aspectos que irão ser apresentados agora, começando pelo aspecto magnitude do efeito (*effect size*) e, posteriormente, analisando o poder dos testes estatisticos.

Magnitude do efeito (*Effect Size*)

Effect size (*ES*), enquanto conceito estatístico, é traduzido normalmente por *tamanho, dimensão* ou *magnitude do efeito* e pode ser definido como o grau em que o fenómeno está presente na população (Cohen, 1988), isto é, diferença efectiva na população. Assim, quanto maior for o *ES*, maior será a manifestação do fenómeno na população. Em termos práticos, o *ES* é uma medida que codifica a informação quantitativa crítica encontrada nos estudos, ou seja, permite estabelecer a diferença real entre grupos (dois, no caso do teste *t*). O *ES* é representado simbolicamente por letras, consoante o tipo de teste. No caso do teste *t* a letra é o «d», na Anova utiliza-se o «f», seguindo a simbologia de Cohen (1988). De modo genérico, poderemos referir que estas medidas vêm de algum modo dar resposta à significância prática dos resultados, seja clínica ou educacional (Conboy, 2003).

Simbolicamente, o *ES* no teste *t* é representado pela letra «d» (Cohen, 1988), mas existem outras letras para o designar, consoante os autores (Hedges e Olkin, 1985; Hunter e Schmidt, 1990). De acordo com Cohen (1988), convenciou-se que os valores de «d» são considerados pequenos se ($.20 \leq d < .50$); médios se ($.50 \leq d < .80$) e grandes se ($d \geq .80$). Este critério constitui-se como uma referência resultante de uma convenção, tendo surgido na década de 60 do século precedente, inspirados a partir dos trabalhos publicados nas áreas da Psicologia e da Educação.

Relembre-se que o valor da estatística do teste *t* para grupos independentes é dado pela equação (e_1), sendo que $gl = n_1 + n_2 - 2$:

$$(e_1)\ t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\left[\frac{(n_1-1)\hat{s}_1^2 + (n_2-1)\hat{s}_2^2}{n_1+n_2-2}\right]\left[\frac{1}{n_1}+\frac{1}{n_2}\right]}}$$

Assim, o cálculo do "d" é dado pela equação (e_2):

$$(e_2)\ d = \frac{|\mu_1 - \mu_2|}{\sigma}$$

É de salientar, ainda, que o teste é bilateral e que os grupos em cada estudo são de igual tamanho, existindo homogeneidade das variâncias, daí que o desvio padrão seja a média aritmética dos desvios padrões das amostras. No Quadro 5 apresentam-se os valores do *ES* calculados para ambos os estudos.

QUADRO 5 – *ES* «d» calculado a partir das estatísticas resumo com recurso à fórmula e_2

Local:	n	$\overline{X}$	$\hat{s}$	*sem*	t	*p*	d (ES)
Urgência da instituição A	70	20.34	3.45	.41	-4.643	.000	.79
Oncologia da instituição A	70	22.94	3.17	.38			
Urgência da instituição B	30	20.65	3.27	.60	-1.903	.062	.49
Oncologia da instituição B	30	22.23	3.16	.58			

Repare-se que, considerando as estatísticas como boas estimativas dos parâmetros respectivos, e de acordo com os valores de Cohen (1988), o valor do *ES* do 1.º estudo (instituição A) é médio (d=.79), apesar de próximo do *ES* grande, enquanto no 2.º estudo o valor é considerado pequeno (d=.49), no entanto, no 2.º estudo as diferenças não se revelaram estatisticamente significativas para α=.05.

O *ES* pode ser transformado em r (coeficiente de correlação ponto bisserial) e *vice-versa*, quer a partir do próprio valor do «d» (equação e_3), quer a partir do *t* do teste e respectivos graus de liberdade (equação e_4), permitindo por esta via obter a variação explicada (VE), tal como o coeficiente de determinação (r^2) na correlação, com uma pequena diferença, os coeficientes do r^2 assumem designações diferentes no caso dos testes, sendo que, no teste *t* se designa por ómega quadrado (equação e_5).

$$(e_3)\ r = \frac{d}{\sqrt{d^2+4}} \qquad (e_4)\ r = \sqrt{\frac{t^2}{t^2+gl}}$$

$$(e_5)\ \omega^2 = \frac{t^2-1}{t^2+(gl+1)}$$

Os valores do ES, do r (coeficiente ponto bisserial), assim como do ω^2 (VE), encontram-se no Quadro 6. Como se pode observar, o resultado do teste referente à comparação das médias do *burnout* dos enfermeiros da instituição A revela que a diferença (das médias) é estatisticamente significativa, com um valor de *ES* muito próximo do elevado (d=.79; r=.370), o que indicia um efeito moderado da natureza do trabalho no *burnout* dos enfermeiros, sendo o valor do ω^2 modesto (13,5%). No que se refere ao estudo efectuado na instituição B, a diferença das médias não se revelou estatisticamente significativa ($t_{(58)}$=-1.903;p=.062), e o valor do *ES* está próximo do médio (*ES*=.49), com r=.24 e ω^2=5,9%.

QUADRO 6 – ES ("d") calculado a partir das estatísticas resumo com recurso à fórmula e_3; e_4; e e_5

Local:	n	$\overline{X}$	$\hat{s}$	**d** (*ES*)	(r)	VE (ω^2)
Urgência da instituição A	70	20.34	3.45	,79	.37	.135
Oncologia da instituição A	70	22.94	3.17			
Urgência da instituição B	30	20.65	3.27	,49	.24	.059
Oncologia da instituição B	30	22.23	3.16			

Note-se que ambos os valores das medidas de tamanho do efeito estão no intervalo considerado por Cohen (1988), como *ES* médio, ainda que, no estudo realizado na instituição B o valor seja de ES=.49, neste caso a diferença não se revelou estatisticamente significativa. Considerando que a amostra foi a maior possível, resta calcular o poder *a posteriori*. Neste sentido vai-se iniciar a apresentação das questões relacionadas com o poder do teste, referenciadas no Quadro 3.

Poder do teste

O poder dos testes é função (Cohen, 1988; Sedlmeier e Gigerenzer, 1989; Murphy e Myors, 1998) do tipo de teste e direccionalidade ou não das hipóteses, da amostra (sensibilidade), do nível de significância e do *ES*.

A sensibilidade de um teste estatístico é, em grande parte, função do tamanho da amostra. Relembre-se que amostras grandes fornecem estimadores mais precisos dos parâmetros populacionais, isto é, à medida que se aumenta o tamanho da amostra,

aumenta a precisão dos estimadores. Os estudos com amostras de reduzida dimensão, e por isso com baixa sensibilidade, transportam uma incerteza considerável quanto à inferência estatística.

Relativamente ao nível de significância utilizada, ele também influencia no poder do teste. Referenciou-se que, convencionalmente, são utilizados os valores de 5% e 1% e algumas vezes 10%. Aumentar o valor do alpha de 1% para 5% ou 10% aumenta o poder do teste. Assim, um alpha de, por exemplo, 5% ou 10% aumenta o poder do teste relativamente à utilização de 1%. No que respeita ao poder do teste e ao *ES*, a relação aponta para que à medida que se aumenta o *ES*, o poder do teste aumenta.

Voltando ao exemplo que foi referido anteriormente, irá proceder-se às análises do poder dos testes. Estas análises são, na maioria das vezes, utilizadas nos desenhos experimentais, mas isso não invalida a sua utilização noutro tipo de desenhos, podendo ser usadas, quer na fase de planeamento dos estudos (*a priori*), por exemplo, para determinar o tamanho da amostra necessário para um determinado poder ou, então, na fase de interpretação dos resultados (*a posteriori*) que é este caso, especialmente, quando as diferenças não são significativas.

O cálculo do poder do teste pode ser efectuado com *software* adequado para o efeito (Faul *et al.*, 2007), dado que, a estimativa manual obriga à utilização das tabelas de Cohen (1988). No quadro 7 são apresentados os valores das medidas de *ES*, assim como, as estimativas de poder (1-β) dos testes, efectuadas com *software*.

QUADRO 7 – Resultados do cálculo das medidas de ES e PO

Local:	n	$\overline{X}$	$\hat{s}$	**d** (*ES*)	r	VE (ω^2)	**PO**
Urgência da instituição A	70	20.34	3.45	,79	.37	.135	.995
Oncologia da instituição A	70	22.94	3.17				
Urgência da instituição B	30	20.65	3.27	,49	.24	.059	.460
Oncologia da instituição B	30	22.23	3.16				

De referir que as estimativas do poder do teste na instituição A mostram que o poder observado (PO) é de 99,5%, enquanto na Instituição B, PO=46%, isto é, tendo como critério para o poder do teste 80%, ou mesmo 50%, de acordo com o critério apresentado por Hill e Hill (2000), é sugerido que a probabilidade de estar a cometer um erro tipo II é elevada.

Neste caso, existindo um *ES* muito próximo do médio, o mais provável é que a não verificação de diferenças se fique a dever ao tamanho da amostra. Então, num novo estudo a realizar, e apontando para um *ES* médio, qual o tamanho de amostra adequado, supondo que se mantém o α=.05? O tamanho da amostra pode ser calculado *a priori* se o investigador tiver os resultados dos estudos efectuados. Neste caso, era necessário, para um poder requerido de 80% (α=.05) e *ES*=.50, uma amostra de 130 enfermeiros, 65 em cada grupo. É naturalmente perceptível porque se refere o efeito do tamanho da amostra sobre os testes de significância.

Conclusão

Quando se refere que algo é *estatisticamente significativo,* tal não quer dizer que seja *significativo* do ponto de vista *clínico* ou *educacional*. Provavelmente esse efeito ou diferença não é "nulo" (na estatística), pode até ser estatisticamente significativo, mas pode ser clinicamente irrelevante, como o contrário também pode ser verdadeiro.

A rejeição da hipótese nula, recorrendo, por exemplo, ao teste *t,* pode evidenciar que existe um efeito significativo e esse efeito pode ser também clinicamente ou educacionalmente significativo. No entanto, amostras grandes (com pequena variação) podem conduzir a um problema, verificando-se a diferença estatisticamente significativa que, todavia, não tem significado efectivo. Também quando se conclui que uma diferença não é significativa, isso não indica propriamente que as médias sejam iguais, ou que não exista um *efeito substantivo*, significa sim que não houve evidência suficientemente forte para *provar* que essas diferenças eram significativas. Se

houve falha (erro tipo II), pode ter sido pelo facto do tamanho da amostra ser inadequado.

Muitas vezes, a apresentação dos resultados dos testes cinge-se apenas aos valores das significâncias obtidas. Repare-se, quando dois estudos sobre a mesma temática, mas em amostras diferentes, produzem valores de *p value* significativos (p=.05 e p=.001), é-se levado a supor que as diferenças são maiores no segundo estudo. Se as amostras forem iguais é bem possível que sim, mas um valor p=.05 num estudo com amostras de reduzida dimensão, pode reflectir um efeito maior do que um p=.001 num estudo que recorreu a uma amostra de grande dimensão. Também, e ainda neste sentido, quando uma dada temática é estudada em duas amostras com a mesma dimensão e características semelhantes, mas o valor encontrado tenha sido de p=.04 numa, e noutra p=.06, isso não significa que apenas no primeiro estudo é que existem diferenças e efeito significativos. O efeito varia de amostra para amostra, assim como o *p value* encontrado.

Relativamente ao *ES*, ele pode ser estimado *a posteriori* da colheita de dados (como para a generalidade dos estimadores) e esse valor pode ser usado na análise do poder do teste. O *ES* é um valor estimado padrão e deve ser o menor efeito que seja importante detectar no caso de uma análise *a priori* para determinar o tamanho mínimo da amostra necessário para determinado poder.

O *ES* não está de todo fora do controlo do investigador. Se um investigador encontra resultados estatisticamente significativos num 1.º estudo, com uma amostra de 30 indivíduos, isso não significa que uma amostra de 30 seja suficiente para planear outro estudo (2.º estudo), pois o valor do *ES* é diferente nos estudos. Poderá, até, acontecer que num 2.º estudo o *ES* seja menor e os resultados estatisticamente não significativos, assim como, no 1.º estudo o valor estimado a partir da amostra pode ser superior ao verdadeiro valor. É relativamente comum falar-se em 30 sujeitos como sendo uma «*grande amostra*» e, por isso, suficiente para realizar um estudo. Se pensarmos no valor do *ES* obtido na maioria dos estudos, o número 30 é desprovido de significado, sobretudo se for para evitar suposições ou pressupostos dos testes paramétricos, como a normalidade das distribuições, entrando-se, já, no campo do equívoco. O cálculo do tamanho da amostra é feito com base numa previsão do *ES*.

Igualmente, o valor de 80% para o poder do teste pode não fazer sentido se não se basear na natureza da temática, tipo de desenho de investigação e, sobretudo, nas consequências lógicas do erro tipo II na decisão do investigador. Se não se encontram diferenças significativas num estudo, a estimação do poder pode ajudar a reflectir sobre os resultados. A terminologia *inconclusivo* refere-se aos casos em que o investigador, mesmo depois de não ter encontrado resultados estatisticamente significativos, considerando o baixo poder e pequeno *ES*, deve manter uma forte suspeita sobre os resultados e pretende questioná-los.

De facto, a decisão em estatística é complexa porque está relacionada com todas as fases do processo de investigação e, principalmente, porque tem implicações no desenho de novos estudos. As revisões sistemáticas, e não apenas as meta-análises como se faz supor, tal como protagonizadas e desejadas actualmente, se não forem efectuadas tendo em linha de conta estas análises e obviamente o seu domínio, correm o risco de decalcar os abstracts ou os quadros e tabelas dos artigos, sem juizo crítico dos estudos produzidos, para além de serem ou não estatisticamente significativos os resultados, com prejuízo para o que se procura evidenciar.

Referências Bibliográficas:

COHEN, J. (1988) - **Statistical power analysis for the behavioral sciences**. 2ª ed. Hillsdale, New Jersey: Lawrence Erbaum.

COHEN, J. (1994) - The earth is round (p < .05). **American Psychologist**. Vol. 49, nº 12, p. 997-1003.

CONBOY, J. (2003) - Algumas medidas típicas univariadas da magnitude do efeito. **Análise Psicológica**. Série 21, nº 2, p. 145-158.

FAUL, F. [et al.] (2007) - G*Power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. **Behavior Research Methods**. Vol. 39, nº 2, p. 175-191.

GLASSER, D. (1999) - The Controversy of significance testing: misconceptions and alternatives. **American Journal of Critical Care**. Vol. 8, nº 5, p. 291-296.

HEDGES, L. ; OLKIN, I. (1985) - **Statistical methods for meta-analysis**. Orlando : Academic Press.

HILL, M.; HILL, A. (2000) - **A investigação por questionário**. Lisboa : Edições Sílabo.

HUNTER, J. ; SCHMIDT, F. (1990) - **Methods of meta-analysis: correcting error and bias in research findings**. Newbury Park : Sage Publications.

JOHNSON, D. (1999) - The insignificance of statistical significance testing. **Journal of Wildlife Management**. Vol. 63, nº 3, p. 763-772.

KENDALL, P. (1999) - Clinical significance [Special section]. **Journal of Consulting and Clinical Psychology**. Vol. 67, p. 283-239.

KERLINGER, F. (1980) - **Metodologia da pesquisa em ciências sociais: um tratamento conceptual**. São Paulo : E.P.V.

MURPHY, K.; MYORS, B. (1998) - **Statistical power analysis: a simple and general model for traditional and modern hypothesis tests**. Mahwah, New Jersey : Erlbaum.

NESTER, M. (1996) - An applied statistician´s creed. **Applied Statistics**. Vol. 54, nº 4, p. 401-410.

RUNYON, R. [et al.] (1996) - **Fundamentals of behavioural statistics**. New York : McGraw-Hill.

SEDLMEIER, P.; GIGERENZER, G. (1989) - Do studies of statistical power have an effect on the power of studies? **Psychological Bulletin**. Vol. 105, nº 2, p. 309-316.

TALLMADGE, G. (1977) - **The joint dissemination review panel ideabook**. Washington: National Institute of Education and the US Office of Education.

WILKINSON, L.; APA TASK FORCE ON STATISTICAL INFERENCE (1999) - Statistical methods in psychology journals: guidelines and explanations. **American Psychologist**. Vol. 54, nº 8, p. 594-604.