UNIVERSITÀ DEGLI STUDI DI CAGLIARI

FACOLTÀ DI STUDI UMANISTICI

CORSO DI LAUREA IN SCIENZE E TECNICHE PSICOLOGICHE

TESI DI LAUREA

La Ricerca in Psicologia:

Tra Significatività e Dimensione dell'Effetto

Relatore: **Studente:**

Dott. Gianmarco Altoè Roberto Merella

ANNO ACCADEMICO 2012/2013

Dedicata ai miei Nonni

INDICE

SOMMARIO

Al giorno d'oggi, dove la scienza si è trasformata sempre di più in un business, è il numero di pubblicazioni a costruire la reputazione di un ricercatore; un criterio efficace, fin tanto che non viene perso di vista lo scopo per cui vengono pubblicate le ricerche. Infatti i testi scientifici hanno il compito di riportare informazioni che siano sempre più accurate, più utili e maggiormente comprensibili per i lettori, in pratica all'evoluzione delle metodologie scientifiche deve corrispondere un'evoluzione delle tecniche di divulgazione delle informazioni.

Il contenuto delle prossime pagine si propone come una revisione della letteratura di lavori nazionali e internazionali in un arco temporale di oltre 40 anni, tutti riguardanti l'approccio di inferenza statistica dominante nelle scienze umane: il *test di verifica dell'ipotesi nulla*.

Il proposito portato avanti si accosta a quello del Manuale di pubblicazione dell'American Psychological Association, cioè fornire alcune delucidazioni e linee guida per la pratica delle ricerche e delle pubblicazioni che intendono avere una valenza scientifica. Ispirandosi all'impostazione dei lavori di F. Agnoli e S. Furlan del 2008 e 2009, questo testo, affronta le criticità dei metodi di inferenza statistica e le loro soluzioni mediante l'utilizzo di quelli che verranno definiti come *indici di dimensione dell'effetto.*

Nel primo capitolo verrà esposta una breve spiegazione circa il metodo di verifica dell'ipotesi nulla, partendo dall'esposizione delle sue origini storiche e concludendo con i punti chiave che ne caratterizzano la procedura.

Il secondo capitolo affronterà la questione delle criticità riguardanti questi test, evidenziando alcune errate interpretazioni riscontrate nella letteratura e alcune debolezze metodologiche dell'approccio.

Infine, nel terzo capitolo, le premesse portate avanti precedentemente saranno usate per descrivere un criterio, suggerito da diversi autori, per sopperire alle carenze dell'approccio di verifica dell'ipotesi nulla.

CAPITOLO 1

L'Approccio NHST

Questo capitolo propone una sintetica spiegazione di cosa sia l'approccio statistico sulla verifica dell'ipotesi nulla, dall'inglese *Null Hypothesis Significance Testing* (NHST). Il capitolo si apre con una introduzione sulle basi storiche che hanno permesso la realizzazione della metodologia NHST e la sua diffusione nelle ricerche scientifiche, e prosegue con una descrizione dei punti fondamentali che compongono il procedimento.

1.1 Introduzione Storica

La statistica inferenziale è una branca della statistica che offre una soluzione al problema dell'impossibilità di compiere misurazioni su una intera popolazione. Questa disciplina racchiude un insieme di tecniche induttive, che hanno lo scopo di giungere a conclusioni affidabili sulla popolazione dall'analisi di un campione tratto da quest'ultima (Welkowitz, Cohen e Ewen, 2006/2009).

La statistica inferenziale può essere suddivisa in quattro settori principali: *teoria della probabilità, teoria dei campioni, teoria della stima* e *teoria della verifica dell'ipotesi.* L'approccio NHST si colloca nell'ultima area e fornisce un insieme di tecniche per determinare, con una certa probabilità, la correttezza delle *ipotesi statistiche*. Con *ipotesi statistica* viene definita "una assunzione (o supposizione) circa le proprietà non note di una o più misurazioni delle popolazioni, tipicamente riguardo i loro parametri o la loro variabilità (distribuzione) tra i valori più piccoli e quelli più grandi" (Bernstein e Bernstein, 1999/2003, pag. 4). Le ipotesi statistiche vengono verificate con il ragionamento deduttivo da quelle che dovrebbero essere le loro conseguenze, vale a dire che si cercano gli eventi che dovrebbero essere veri se fosse vera l'ipotesi.

Nell'articolo di Agnoli e Furlan (2008) è presente un resoconto storico molto efficace che verrà ripreso per alcuni tratti. Tra il diciottesimo e il ventesimo secolo, in diversi contesti, sono state utilizzate cinque applicazioni dei test statistici (Huberty,

1993). Malgrado la precisione con cui queste ultime sono state rintracciate, le origini dell'approccio di verifica dell'ipotesi nulla (NHST) non sono chiare, ciononostante si sa che si è diffuso in America e successivamente in Europa tra il 1940 e il 1955. Dagli anni '60 i test sulla significatività statistica diventarono dominanti nelle ricerche delle discipline scientifiche come la psicologia, la medicina o la sociologia; ben l'80% degli articoli pubblicati nelle riviste internazionali contenevano procedure collegabili a questi approcci (Agnoli e Furlan, 2008).

A discapito della sua enorme diffusione, già dagli anni '30 la verifica dell'ipotesi nulla è stata ampiamente criticata, anche nel mondo della psicologia sperimentale: Skinner fondò il *Journal of Experimental Analysis of Behavior* per sfuggire alle dinamiche dell'NHST (Luccio, Salvadori e Bachman, 2005). Nel corso del tempo il dibattito ha coinvolto numerosi autori, alcuni si sono espressi criticando l'errato uso del metodo (Cohen, 1994; Gigerenzer, Krauss e Vitouch, 2004), altri invece ne hanno messo in dubbio il valore stesso raccomandandone addirittura l'abolizione (Roseboom, 1960; Schmidt, 1996).

L'NHST è una procedura nata dalla fusione di due diversi approcci: il primo è detto *p-value Approach* (PVA); il secondo è chiamato *Fixed Alpha Approach* (FAA). Per poter capire appieno i fraintendimenti che verranno esposti nel capitolo successivo, come quelli legati al senso del *p-value* o della *significatività statistica*, è necessario spiegare, seppur sinteticamente, le caratteristiche dei modelli sopra citati.

Il PVA, ideato da R.A. Fisher, è stato descritto per la prima volta nel suo libro *Statistical Method for Research Workers* del 1925 e successivamente rivisto nel corso degli anni. Il procedimento può essere definito come un *test sulla significatività*, si basa sul presupposto che partendo dalle caratteristiche note di una popolazione si possano formulare delle previsioni sulle caratteristiche che dovrebbe avere un campione estratto da essa; nel momento in cui venissero violate queste aspettative, ad esempio attraverso una *significativa* differenza tra l'andamento noto del parametro e quello derivante dai dati, allora si può inferire che il campione è stato tratto da una popolazione diversa da quella conosciuta (Luccio et al., 2005).

Riprendendo l'articolo di Hubbard e Bayarri (2003) si può dire che: definita un ipotesi (H) e raccolti dei dati (x), il PVA usa le discrepanze nei dati per rifiutare l'ipotesi, perciò la probabilità associata ai dati (*p-value*) fornisce la veridicità di H,

oppure P (x | H). L'ipotesi viene *rifiutata* quando la probabilità calcolata è inferiore a un criterio chiamato *Livello di Significatività*. Secondo le parole dello stesso Fisher (1966, come riportato da Hubbard e Bayarri, 2003, pag. 3) "è comune e conveniente, per le sperimentazioni, prendere il 5% come livello standard di significatività*"*.

Più avanti in questo lavoro verrà evidenziato come il suggerimento fishieriano sopracitato ha assunto un'importanza crescente e problematica all'interno dell'inferenza statistica, nonostante il fatto che lo stesso Fisher (come citato da Gigerenzer, 1994) nei suoi ultimi scritti abbia suggerito di adattare il livello di significatività ai casi particolari, poiché impensabile applicare una metodologia universalmente valida.

Per quanto riguarda il secondo approccio Luccio et al. (2005) spiegano che il FAA è stato descritto nel 1933 da J. Neyman e E.S. Pearson, i quali lavorarono per anni ai problemi più salienti nel dibattito statistico del tempo. A differenza di Fisher gli autori impostano il modello sulle tecniche per prendere una decisione tra due ipotesi esaustive e mutualmente escludenti, indicate con H_0 e H_1.

Nel prossimo capitolo verrà esposta l'importanza di alcuni concetti che derivano da questa teorizzazione: l'errato rifiuto di H_0 quando è vera, detto *errore di I tipo* (indicato con α), l'errata accettazione di H_0 quando è falsa, detta *errore di II tipo* (indicato con β) ed infine la *potenza statistica*, definita come la probabilità di rifiutare correttamente H_0 quando è vera H_1, uguale a $(1 - \beta)$.

Nel FAA, viene fissato a priori il livello di α (da qui il nome del modello) e grazie a questo si calcola la zona di rigetto, cioè l'insieme dei valori del parametro studiato dai dati che causano il rifiuto di H_0. Dopodiché si cerca il test con la potenza più alta per diminuire la probabilità di commettere *errori* (Christensen, 2005). Huberty (1993) fa notare che il FAA può essere definito come un *test sulle ipotesi*, per differenziarlo dal PVA. Il modello di Neyman e Pearson prende in considerazione una *ipotesi nulla* H_0, una *ipotesi alternativa* H_1 e una *regione di rigetto* calcolata a partire da un valore di probabilità α fissato a priori. Dopo aver trovato il valore della statistica campionaria, se questa cade nella regione di rifiuto allora si esclude H_0 in favore di H_1, in caso contrario viene tollerata.

Anche da una trattazione storica così ridotta emergono le grosse differenze che separano i due approcci; nonostante questo, il modello attualmente impiegato

nelle ricerche psicologiche, consiste in un ibrido che cerca il livello di significatività (tipico del modello fisheriano) all'interno di un confronto tra ipotesi statistiche (ripreso chiaramente da Neyman e Pearson).

Il prossimo paragrafo descriverà più specificatamente il *Null Hypothesis Significance Testing* con lo scopo di evidenziare come l'incrocio tra le due metodologie descritte precedentemente abbia dato forma ad una struttura che, se applicata da sola, risulta incompleta e ricca di ambiguità.

1.2 Il metodo

All'inizio del paragrafo precedente è stato detto che l'NHST si colloca all'interno della statistica inferenziale e per questa ragione il suo obiettivo è quello di verificare delle ipotesi su una popolazione non misurabile, per fare ciò utilizza dei dati ottenuti da un campione misurabile; pertanto, la domanda a cui cerca di rispondere può essere espressa in questo modo: come si può giudicare la veridicità di una ipotesi riguardo una popolazione?

Data una generica statistica t riguardo la popolazione che possiede un andamento noto se fosse vera l'ipotesi nulla, la verifica avviene traslando i dati del campione (x) nell'andamento teorico di t e valutando la probabilità che hanno di appartenere a tale distribuzione. Nel caso in cui sia sufficientemente bassa si dice che la statistica campionaria t(x) devia in modo *significativo* dalla distribuzione sotto l'ipotesi nulla. Riferita al contesto statistico (poiché più avanti verranno fatte delle distinzioni riguardo altri contesti), con il termine *significatività* viene indicato il fatto che è stata registrata una deviazione del campione dall'andamento della popolazione tale che la probabilità associata ad essa è talmente bassa da considerare l'ipotesi nulla improbabilmente vera.

Agnoli e Furlan (2009), hanno riassunto il ragionamento che sta dietro l'NHST nei seguenti passaggi:

1. Se l'ipotesi H fosse vera allora è altamente improbabile che il test sia significativo
2. Il test è significativo
3. Allora H è altamente improbabile

Sul piano della logica formale, la verifica dell'ipotesi nulla si basa sull'implicazione "se P allora Q", da cui si può dedurre "se Q è falso allora P è falso". Questo procedimento però può andare solo in direzione falsificazionista, poiché non è possibile affermare che "se Q è vero allora P è vero" (Bakan, 1966).

Nei prossimi paragrafi verrà esposta la procedura di verifica dell'ipotesi nulla con un attenzione particolare verso la significatività statistica, per la quale viene ripreso il concetto fisheriano di *p*.

1.2.1 Null Hyphotesis Significance Testing

Per fornire una spiegazione esaustiva del metodo è possibile rifarsi a diversi manuali di statistica inferenziale, questo lavoro considera principalmente la versione italiana del testo di S. Bernstein e R. Bernstein (1999/2003).

In ogni studio in cui si vuole applicare la verifica dell'ipotesi nulla il primo problema di un ricercatore è quello di estrarre un campione che sia il più possibile rappresentativo della popolazione che si intende studiare. A questo scopo vengono utilizzate delle tecniche dette di *campionamento*, le quali non verranno affrontate in questo lavoro e proseguiremo dando per scontato che la selezione del campione e la raccolta dei dati sia avvenuta in maniera ottimale. Nell'NHST vengono formulate due ipotesi in contrasto tra loro: la prima, che viene considerata vera nella popolazione oggetto di studio, è detta *ipotesi nulla* e viene indicata con H_0, la seconda, invece, è detta *ipotesi alternativa* (o "del ricercatore") e viene indicata con H_1.

L'ipotesi H_0 è detta "nulla" o "a differenza nulla" perché suppone l'assenza di una differenza tra il parametro ignoto μ; della popolazione e il parametro campionario μ_0, perciò nella sua più tipica formulazione risulta:

$H_0 : \mu = \mu_0$

L'ipotesi H_1 invece è detta "di ricerca" o "del ricercatore" ed afferma che il parametro campionario differisca da μ, essa può essere formulata in diversi modi:

1. $H_1 : \mu \neq \mu_0$

   In questo caso viene detta *ipotesi alternativa bidirezionale,* poiché indaga tutte le possibili differenze in entrambe le direzioni.

2. $H_1 : \mu > \mu_0$ oppure $H_1 : \mu < \mu_0$

In questi casi viene definita *ipotesi alternativa unidirezionale* perché suppone delle differenze con il parametro incognito in una sola direzione.

Come si può facilmente intuire, qualunque modalità si scelga per H_1, le ipotesi formulate sono complementari ed esaustive; vale a dire che, insieme, coprono tutta la gamma di opzioni possibili, motivo per cui H_1 viene considerata vera se H_0 risultasse significativamente improbabile.

Come secondo passaggio, dopo la formulazione delle ipotesi, si calcola una statistica dei dati *x* chiamata *statistica test* ed indicata con t, tale per cui t = t (*x*). La statistica scelta ha la particolarità di avere un andamento noto se fosse vera H_0, quindi $f(t \mid H_0)$, e perciò è possibile calcolare la probabilità di ottenere un risultato uguale o più estremo rispetto a quello osservato empiricamente (questa probabilità è il *p-value* di cui si è accennato precedentemente).

Ogni test si conclude con una presa di posizione riguardo H_0, che può essere tollerata o rifiutata. Riprendendo concezioni tipiche del FAA, con questa decisione si possono commettere due tipi di errori: rifiutare H_0 quando è vera (cioè α) e tollerare H_0 quando è falsa (cioè β). A questo punto però emerge un problema: come si può stabilire se il *p-value* è abbastanza piccolo per poter rifiutare H_0?

Come soluzione i ricercatori hanno utilizzato per anni il "suggerimento" di Fisher e hanno assegnato ad α un valore di probabilità convenzionale di .05 per la maggior parte degli studi oppure di .01, o addirittura di .001, per quelli più restrittivi. Quando $p \leq \alpha$ allora l'ipotesi nulla viene rifiutata con un livello di probabilità considerato accettabile.

Detto ciò, è opportuno precisare alcune questioni: in nessun caso è consigliabile parlare di "accettazione" di una o dell'altra ipotesi. La verifica dell'ipotesi nulla permette di prendere decisioni probabilistiche riguardo il rifiuto dell'ipotesi nulla, ciononostante non comunica nulla riguardo la veridicità delle ipotesi formulate. Il livello di α è deciso prima di fare l'esperimento, mentre quello di β è destinato a rimanere ignoto se non si conosce il reale risultato della popolazione, rendendo sconosciuta la probabilità di commettere l'errore del II tipo. É dimostrato che esiste una relazione inversa che lega il livello dell'errore di I tipo e quello dell'errore di II tipo, ciò significa che l'interesse dei ricercatori nel ridurre a priori α ha come conseguenza un aumento di β, un problema che verrà spiegato meglio nel

prossimo capitolo insieme al concetto di *potenza statistica*. In aggiunta, l'NHST è vittima della paradossale situazione per cui il ricercatore non ottiene nessuna informazione riguardo H_1 (ipotesi del ricercatore), poiché l'intera procedura si basa esclusivamente sull'analisi dell'ipotesi nulla.

L'insieme degli elementi appena proposti rendono esplicita l'impossibilità di avere una certezza assoluta sulla posizione che il ricercatore prende nei confronti di una ipotesi e, poiché la verifica delle ipotesi si basa su un ragionamento di falsificazione, è consigliabile non accettare alcuna ipotesi, ma limitarsi a *rifiutare* H_0 oppure *non rifiutare* H_0.

1.2.2 Il p-value

Come si può notare dalla spiegazione precedente, nella verifica dell'ipotesi nulla il concetto cardine è senza dubbio quello di *p-value.* Per chiarire questo costrutto è necessario fornire alcune premesse: esistono diverse tipologie di statistiche test che possono essere utilizzate per la verifica della significatività statistica, considerando vera a priori l'ipotesi che la popolazione abbia una specifica caratteristica allora ognuna di queste statistiche ha un andamento noto, questo andamento teorico rappresenta l'insieme dei possibili risultati della statistica campionaria $t(x^{rep})$ che si otterrebbero se l'esperimento venisse ripetuto infinite volte.

In seguito alla raccolta dei dati sperimentali si osserva dove la t(x) acquisita dalle misurazioni si colloca nella distribuzione teorica. Quanto detto è reso più facilmente comprensibile grazie allo schema proposto da Wagenmakers nel 2007 (vedi Fig. 1.1).

Come descritto nel recente articolo di Pastore (2009), il *p-value* è una probabilità condizionata; geometricamente calcolare l'area sottostante la distribuzione nota*,* chiamata anche funzione di densità, da un determinato punto t(x) fornisce la probabilità di ottenere un valore uguale o più estremo a quello considerato. Questa probabilità non è altro che il *p-value* e può essere descritto dalla formula:

$$p\text{-}value = \text{Prob}\,[\,|\,t\,|\,(\,x\,|\,H_0\,) \geq |\,t\,|\,(x)\,]$$

ovvero la probabilità di osservare un valore maggiore della statistica campionaria a

condizione che H_0 sia vera (Pastore, 2009).

La formula prende in considerazione il valore assoluto della statistica campionaria t(x), perché nel caso di una ipotesi alternativa bidirezionale si valuta la deviazione su entrambe le code della distribuzione nota (come mostra la parte tratteggiata del grafico nella Fig.1.1).

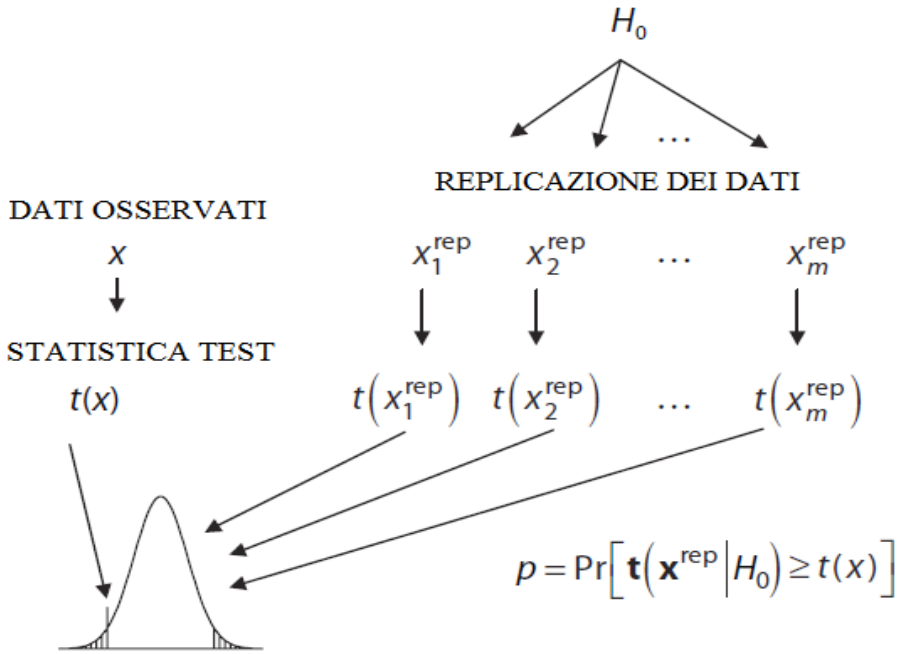

Fig. 1.1 Schema di rappresentazione grafica del *p-value* (Wagenmakers, 2007)

1.2.3 Aspetti critici

Quanto esposto fino a qui è una sintesi dell'approccio NHST, per questa ragione è bene precisare che ci sono alcuni aspetti che hanno generato incomprensioni tra i ricercatori che si sono affidati alla verifica dell'ipotesi nulla. Il fattore principale è il criterio con il quale viene deciso se sia opportuno rifiutare H_0; più precisamente, vengono spesso mal interpretati i concetti di *p-value* e di risultato *statisticamente significativo.*

Pubblicare una ricerca che ha prodotto dei risultati statisticamente significativi comunica solamente che è stato calcolato un valore di *p* inferiore al valore fissato di α, pertanto viene rifiutata H_0 in base all'idea che ci sia una

probabilità troppo bassa per considerare vera l'ipotesi di aver estratto un campione da una popolazione in cui valga H_0. Nel prossimo capitolo, invece, verranno affrontate alcune diffuse interpretazioni erronee riguardanti la significatività statistica.

Dalla definizione però appare chiaro che tale pubblicazione non fornisce alcuna informazione né sulla forza dell'effetto che è oggetto di studio né sulla qualità dell'esperimento; ad esempio, una differenza tra medie statisticamente significativa dice solamente che è improbabile aver estratto dalla stessa popolazione i due campioni presi in analisi (Nickerson, 2000). Per via della mancanza di informazioni sulla grandezza dell'effetto, o se questo sia di una qualche utilità pratica, è sconsigliato utilizzare esclusivamente la significatività statistica come criterio di selezione delle ricerche.

Oltre alle errate interpretazioni, gli aspetti controversi del *p-value* coinvolgono vere e proprie criticità concettuali. Ai fini di questo lavoro ne verranno fatte presenti solo alcune:

- **il *p* dipende dalle intenzioni del ricercatore:** ciò si riferisce al fatto che i valori del *p-value* dipendono dalle tecniche di campionamento utilizzate. Il problema sorge nel momento in cui i ricercatori ignorano questa influenza, dato che le tecniche di campionamento vengono scelte secondo previsioni soggettive di eventi futuri, non c'è alcun modo per controllare se i *p-value* ottenuti siano di parte o meno (Wagenmakers, 2007).
- **il *p* dipende dalla numerosità campionaria:** il calcolo delle statistiche test maggiormente diffuse (ad esempio *F*, *t*, chi-quadro) tendono a dare valori di *p* minori all'aumentare della numerosità campionaria, da ciò ne consegue che per un ricercatore sia più facile raggiungere la significatività statistica se dispone di un ampio campione (Daniel, 1998).
- **il *p* è una misura dell'evidenza contro H_0:** poiché valuta la probabilità di osservare i dati a condizione che valga l'ipotesi nulla, cioè $p\,[\, t(x) \mid H_0 \,]$. Con la verifica dell'ipotesi nulla basata unicamente sul calcolo di *p*, non è possibile sostenere H_0, ma solo evidenziare se è falsa o, quanto meno, se la si può considerare falsa con un margine di errore accettabile.

La probabilità che l'ipotesi nulla sia vera a partire dai dati osservati, cioè P[H_0 | t(x)], è propria di un metodo parallelo all'NHST: l'approccio bayesiano. Questo lavoro non entrerà nel merito di questo approccio, tuttavia è possibile trovare alcune informazioni al riguardo nel recente articolo di M. Pastore (2009).

CAPITOLO 2

I Limiti dell'NHST

In questo capitolo verrà affrontato il tema delle debolezze che, nel corso degli anni, diversi autori hanno evidenziato nel Null Hypothesis Significance Testing e nel suo utilizzo. Il capitolo si apre con un resoconto su alcune interpretazioni erronee presenti nella letteratura scientifica, successivamente è presente una spiegazione del concetto di *potenza statistica* e del suo ruolo all'interno delle ricerche, infine, il capitolo si chiude con una nota riguardo la differenza tra significatività statistica e importanza pratica, la quale ha la funzione di introdurre il tema conclusivo di questo lavoro.

2.1 Problemi di interpretazione

Il resoconto storico fatto nel primo capitolo conteneva un accenno alle critiche che sono state mosse contro l'approccio di verifica dell'ipotesi nulla. Il dibattito si è prolungato nel tempo e ha coinvolto una quantità enorme di ricercatori. Per porre fine alle polemiche l'*American Psychological Association* (APA) istituì una commissione, la *Task Force on Statistical Inference* (TFSI), con lo scopo di studiare le caratteristiche dei problemi portati alla luce e cercare eventuali soluzioni. I diversi rapporti della TFSI diedero il loro contributo nelle revisioni che il Manuale APA ha avuto fino ad oggi; nella sesta edizione, uscita nel 2010 (mentre la versione italiana è uscita nel 2011), vengono elencate l'insieme di norme per un corretto procedimento di ricerca e per una facile comprensione della scrittura scientifica (Vellone e Alvaro, 2011).

All'interno del rapporto pubblicato nel 1999 da Wilkinson e la TFSI ci sono alcuni suggerimenti rivolti ai ricercatori, tra i quali: 1) definire chiaramente il tipo di studio che si intende condurre, se la ricerca possiede più obiettivi è necessario assicurarsi che siano definiti e che venga resa nota la loro priorità, 2) rendere note tutte le caratteristiche essenziali per l'esperimento (come le modalità di campionamento, la popolazione considerata), poiché l'interpretazione dei risultati

dipende dalle caratteristiche della popolazione in analisi, 3) rinunciare alla scelta dicotomica tra rifiutare o meno H_0 e riportare semplicemente il valore esatto di *p,* meglio se con degli intervalli di confidenza e, 4), accompagnato dai valori della *potenza statistica* e da stime sulla *dimensione dell'effetto* (Wilkinson e TFSI, 1999).

L'ampio dibattito che ruota attorno all'NHST ha visto, dalla parte delle critiche, sia autori che ne hanno esposto aspramente le debolezze a favore di altri metodi di indagine statistica, come Roseboom (1960) o Schmidt (1996), sia autori come Gigerenzer et al. (2004), che nonostante abbiano sottolineato le debolezze dell'approccio, consigliano delle strategie metodologiche per compensarne i punti deboli, prima tra tutte un uso più consapevole degli strumenti che si intendono utilizzare. Sul versante opposto, quello dei favorevoli all'NHST, si trovano autori come Wilkinson con la TFSI e Abelson (1997). Quest'ultimo, infatti, suggerisce alcuni criteri che devono essere considerati per realizzare una buona ricerca e sostiene che rinunciare ai test di significatività eliminerebbe un importante strumento di selezione delle ricerche, un "portiere", che seleziona cosa può essere pubblicato in base alla qualità del lavoro.

Qualunque posizione si intenda prendere nei confronti dell'NHST, è indubbio che anche attualmente è presente una notevole confusione riguardo le indagini statistiche; una tra le principali problematiche è l'uso meccanico ed irragionevole che questo approccio ha subito: attraverso una "ritualizzazione" la verifica dell'ipotesi nulla ha acquistato una valenza assoluta che molti ricercatori hanno smesso di mettere in discussione, talvolta anche causando un rallentamento del progresso scientifico (Fidler, 2005). Gigerenzer et al. (2004) paragonano la verifica dell'ipotesi nulla ad un "martello", che i ricercatori utilizzano per qualsiasi occasione, non facendo caso all'ampia gamma di "attrezzi" che sono a loro disposizione.

Tra le procedure maggiormente ritualizzate l'impostazione di α ad una probabilità dello .05 è una delle più controverse. Questa convenzione ha portato molti ricercatori a prendere una decisione dicotomica esclusivamente in base al valore di *p*: se minore di α allora si pubblica la ricerca come significativa, in caso contrario non la si considera. Frick (1996), parla del valore .05 come di una "scogliera", al di sotto del quale, o al di sopra della quale, tutti i valori di *p* vengono trattati con la medesima importanza. Valori statisticamente significativi come $p = .04$

e p = .001 vengono considerati come ugualmente importanti e saranno pubblicati, mentre sia p = .80 che p = .06 subiscono un trattamento completamente opposto, nonostante né p = .04 né p = .06 offrano argomentazioni schiaccianti a favore o contro l'ipotesi nulla (Frick, 1996). Utilizzare l'NHST in maniera prettamente meccanica potrebbe condurre alla paradossale situazione in cui due ricercatori traggono conclusioni diverse nonostante abbiano rilevato gli stessi effetti, unicamente perché hanno utilizzato campioni con numerosità differenti. Perciò, il ricercatore che ha il campione più piccolo rileva un p = .06 e non può rifiutare l'ipotesi nulla; mentre dall'esperimento realizzato con un campione più grande si ottiene p = .04 e sarà pubblicato come statisticamente significativo (Agnoli e Furlan, 2009).

Rosnow e Rosenthal (1989) scrivono del *p-value* sostenendo che, con il tempo, abbia acquistato una sorta di "ontologia mistica", al quale i ricercatori riconoscono effetti positivi esclusivamente al di sotto della soglia di significatività statistica. Il livello di α, che sia impostato a .05 oppure a .001, è un limite convenzionale, e non uno ontologico, tra ciò che è significativo e ciò che non lo è.

Nel manuale di Bernstein e Bernstein (1999/2003), vengono descritti i punti di vista di un *produttore* e di un *consumatore*. Il primo ragiona come un tipico ricercatore ed è interessato a ridurre al minimo la probabilità che i lotti da lui prodotti vengano rifiutati nonostante siano conformi ai requisiti del consumatore; ciò significa che il "rischio del produttore" non è altro che l'errore di I tipo, ovvero rifiutare H_0 quando è vera. Al contrario, il consumatore ha una visione opposta: per lui è prioritario evitare di comprare un lotto che non rispetta i requisiti, di conseguenza sarà intenzionato a diminuire l'errore del II tipo, ovvero tollerare H_0 quando è falsa. Con questa descrizione, emerge abbastanza chiaramente come l'obbiettivo specifico di un ricercatore possa influire sulla progettazione di una sperimentazione; affinché venga massimizzata l'efficienza degli strumenti statistici è una buona pratica aggiustare i livelli di α in base ai casi specifici.

Le interpretazioni erronee però non si limitano ai concetti procedurali, ma anche a quelli descrittivi. Un'altra questione con numerosi fraintendimenti riguarda infatti il significato dell'etichetta "statisticamente significativo", la quale comunica che è stata calcolata una probabilità troppo bassa per considerare vera l'ipotesi di aver

estratto il campione da una popolazione in cui vale H_0. Carver (1978) parla di "fantasie sulla significatività statistica" e le divide in tre categorie principali: (a) credere che .05 sia la probabilità che i risultati siano dovuti al caso, oppure che non lo siano con una probabilità di .95, (b) credere che ci sia una probabilità di .95 che i risultati possano essere replicati (questo errore è chiamato *replicability fallacy*), e (c) credere che .95 sia la probabilità che l'ipotesi sia vera. A queste categorie potrebbero essere aggiunte (d) la convinzione che la significatività statistica indichi la rappresentatività del campione rispetto alla popolazione oppure (e) la validità dell'esperimento (Daniel, 1998).

Su questo argomento Haller e Krauss (2002) hanno condotto una ricerca per indagare le misconcezioni di studenti e insegnanti nei corsi di psicologia riguardo ai concetti di *p* e di *significatività statistica.* Lo studio prevedeva la somministrazione di un questionario con item a risposta chiusa, nel quale si doveva indicare se gli enunciati fossero veri o falsi; questi ultimi rappresentavano i tre tipi di misconcezioni più comuni: 1) classificare un risultato significativo come una prova assoluta a favore dell'ipotesi nulla o di quella alternativa, 2) considerare il *p-value* come la probabilità bayesiana che l'ipotesi nulla sia vera a partire dai dati, cioè $P [H_0 | t(x)]$, 3) ritenere che il complemento di *p*, cioè $1 - p$, corrisponda alla probabilità di osservare nuovamente lo stesso risultato se si tenessero costanti le condizioni sperimentali.

I risultati furono molto interessanti e mostrarono che gli errori più comuni riguardarono i punti 2 e 3. Per questa ragione gli autori suggerirono che l'insegnamento della statistica agli studenti di psicologia non dovrebbe limitarsi a spiegare formule e procedure, ma dovrebbe concentrarsi sul diffondere la comprensione e il "pensiero statistico". A questo proposito elencarono una serie di quattro "passi" per comprendere il senso della verifica dell'ipotesi nulla:

1. insegnare che esistono due diversi paradigmi statistici, l'NHST e l'approccio Bayesiano
2. chiarire che l'NHST considera $p [t(x) | H_0]$
3. definire che l'approccio Bayesiano valuta $P [H_0 | t(x)]$
4. stimolare gli studenti al confronto tra i due metodi

Esiste anche un'altra considerazione da fare riguardo la significatività statistica: riferirsi ad un risultato con il termine "significativo" è alla base di malintesi linguistici. Se utilizzato in ambito statistico ha infatti una accezione diversa da quella del senso comune, tuttavia nella parte conclusiva di alcuni articoli, appaiono commenti come "altamente significativo" o "importante", etichette fuorvianti oltre che erronee (Cohen, 1994).

Dopo aver preso in analisi alcuni dei più importanti equivoci intorno all'NHST il prossimo paragrafo affronterà una tematica cruciale per la validità di una ricerca: la *potenza statistica.*

2.2 La potenza statistica

La potenza statistica può essere definita come la probabilità di rifiutare H_0 quando essa è falsa e matematicamente viene descritta come il complemento di β, cioè 1 – β. Come accennato nel capitolo precedente, non è possibile determinare il valore di β senza conoscere il valore reale del parametro μ della popolazione; tuttavia è possibile calcolare una distribuzione di valori possibili di β a partire da un insieme di valori alternativi di μ (Bernstein e Bernstein, 1999/2003). Il calcolo della potenza è influenzato da fattori come la numerosità campionaria, il livello di significatività, la direzionalità di H_1 e il test statistico scelto.

Come accennato nel primo capitolo, il concetto di *potenza* deriva dall'approccio di Neyman e Pearson, tuttavia a causa della fusione con il metodo fisheriano e della ritualizzazione in cui è caduto l'NHST il calcolo e l'analisi della potenza hanno finito per essere rapidamente trascurati.

L'*analisi della potenza* è una procedura consigliata sia per la progettazione che per la valutazione di una ricerca; l'articolo di Cohen (1992a) spiega che essa prende in considerazione la relazione matematica tra quattro variabili: numerosità campionaria, dimensione dell'effetto, α e potenza. Il loro rapporto permette di determinare uno qualsiasi dei fattori quando si conoscono gli altri tre. Attualmente vengono impiegate due forme di analisi della potenza: la prima permette di stabilire a priori la numerosità campionaria necessaria per raggiungere la potenza desiderata, a partire da una dimensione dell'effetto teorica; la seconda invece calcola la potenza da

una numerosità ed un α stabiliti, per individuare una dimensione dell'effetto ipotizzata (questa forma di analisi è adoperata nelle meta-analisi).

Nello stesso articolo, Cohen (1992a) scrive che, in assenza di qualsiasi base per l'impostazione della potenza a partire da conoscenze pregresse, considerare un valore di .80 potrebbe essere un convenzionale livello di potenza desiderabile. Ancora una volta l'accento va messo alla parola "convenzionale", affinché non venga interpretato come la "scogliera" del *p-value*.

La potenza statistica è influenzata dalla dimensione dell'effetto che si desidera cercare e dalla numerosità campionaria. Nella critica portata avanti da Schmidt (1996) si parla di una potenza dello .80 come standard utilizzato nella ricerca psicologica; tuttavia è possibile raggiungere un livello di potenza simile principalmente in ricerche che si occupano di studiare effetti "grandi". Una indagine che si occupa di confrontare la differenza tra due tipi di trattamento potrebbe avere un effetto talmente basso che per ottenere la potenza desiderata sarebbe necessario un campione eccessivamente numeroso.

Quanto detto però non implica che per effetti piccoli non sia possibile compiere ricerche, e persino uno studio con una bassa potenza potrebbe dare il suo contributo alla conoscenza scientifica in un confronto nelle meta-analisi (Schmidt, 1996). Ciò che deve essere evitato è l'esclusione della probabilità di β, e quindi della potenza del test, nella costruzione di una ricerca. Nel libro di Fiona Fidler (2005) è presente un efficace resoconto delle conseguenze negative che alcune pubblicazioni, non riportando analisi della potenza, hanno provocato alla conoscenza scientifica, anche solamente ritardando per anni l'approfondimento di alcune tematiche.

La potenza bassa di un test non è nociva di per se ma deve sempre essere interpretata con occhio critico da parte del ricercatore. Considerare inutile a priori uno studio con un livello di potenza basso è un errore basato sull'idea che ogni pubblicazione vada valutata con un test sulla significatività statistica e debba poter supportare una conclusione (Schmidt, 1996). L'invito che viene fatto è sempre lo stesso, una maggiore consapevolezza da parte di chi intende condurre delle sperimentazioni. A questo proposito la Fidler (2005) scrive:

Purtroppo, la bassa potenza statistica non è l'unico problema. In realtà, non è

> nemmeno il problema principale, in quanto le meta-analisi offrono l'opportunità per gli studi, con bassa potenza, di dare un importante contributo alla letteratura di ricerca. Spesso non c'è modo per evitare un esperimento di bassa potenza. Ad esempio, quando si lavora con gruppi naturali, non si può infettare pazienti in più con una malattia rara, o aumentare la popolazione di una specie in pericolo. Condurre uno studio su larga scala può essere semplicemente troppo costoso o richiedere tempo. Tutte queste cose sono comprensibili e giustificabili. Gravi problemi sorgono solo quando la potenza statistica è bassa *e* sconosciuta. (p. 43)

Si può notare che la preoccupazione principale dell'autrice è la mancata comunicazione della potenza statistica. La stessa TFSI dell'APA (1999) ne ha raccomandato l'utilizzo per delle pubblicazioni più trasparenti e facilmente comprensibili.

Questo lavoro, fino ad ora si è soffermato sul corretto svolgimento di una ricerca eseguita utilizzando l'NHST; il prossimo paragrafo, invece, affronta una delle tematiche più salienti e più controverse: la differenza tra significatività statistica e importanza pratica.

2.3 Significatività statistica contro importanza pratica

Kirk (2001) scrive che le domande a cui un ricercatore tenta di rispondere sono tre: l'effetto osservato può essere considerato vero o è attribuibile al caso? Se è reale, quanto è grande questo effetto? È di una qualche utilità? La prima domanda è chiaramente competenza dell'NHST e può trovare risposta attraverso la significatività statistica. La seconda riguarda la grandezza dell'effetto e verrà trattata più dettagliatamente nel prossimo capitolo. L'ultimo quesito, invece, riguarda la differenza presente tra la significatività statistica e l'importanza pratica.

Nel capitolo precedente è stato precisato che cosa comunica l'espressione “statisticamente significativo” nel commentare il risultato di una ricerca. Tuttavia esistono diverse tipologie di significatività che i ricercatori dovrebbero tenere in considerazione: statistica, pratica, clinica (Thompson, 2002). Questo fatto appare più evidente nelle occasioni in cui i diversi tipi di significatività non coincidono.

Una pubblicazione con risultati statisticamente significativi implica solamente che l'ipotesi nulla è stata rifiutata, non fornisce informazioni sulla forza dell'effetto o sulla sua importanza pratica. Quest'ultima è influenzata dalla qualità del disegno di ricerca e dalla rilevanza che ha l'oggetto di studio. Affinché dei risultati siano importanti devono avere a che fare con problemi importanti (Huck, 2012). Malgrado ciò, attualmente viene dato maggior peso alla significatività statistica come fattore per stabilire quali ricerche siano degne di nota.

Questo stile decisionale alimenta la falsa credenza che al diminuire dei valori del *p-value* corrisponda un aumento dell'importanza pratica (Kraemer et al., 2003), un interpretazione che sarebbe valida esclusivamente se i fattori che influenzano *p* (Vedi paragrafo 1.2.3) rimanessero costanti. Dato che la significatività statistica è in funzione del *p-value* allora tutto ciò che comporta la variazione di uno porterà inevitabilmente una variazione dell'altra.

Nel capitolo precedente è stato accennato che solitamente, a parità di effetto, l'aumento del campione comporta una diminuzione dei valori di *p*. Perciò, con un campione sufficientemente grande sarebbe possibile ottenere un risultato statisticamente significativo, che tuttavia non possiede alcuna rilevanza pratica. Agnoli e Furlan (2008) descrivono efficacemente questo fenomeno riprendendo G. Loftus (2002); quest'ultimo ipotizza un'indagine condotta su un campione di 10,000 soggetti per condizione sperimentale che ha dato un risultato statisticamente significativo. Lo studio consisteva su una analisi della relazione tra l'assunzione quotidiana di un certo dosaggio di vitamina C e la media dei giorni passati con il raffreddore; nella Tabella 2.1 sono riportati i dati sperimentali.

Tra il gruppo con il dosaggio di 2 grammi e quello da 4 grammi è presente una differenza in media di .23 giorni. Malgrado ciò è possibile ottenere un risultato significativo in termini statistici grazie all'elevata numerosità campionaria, ma privo di qualsiasi rilevanza se si considera la media effettiva dei giorni trascorsi con il raffreddore.

A questo punto, un ricercatore che prende coscienza della possibile discordanza tra l'importanza che gli stessi risultati hanno in senso statistico o in senso pratico, deve confrontarsi con il metodo utilizzato per stabilire la rilevanza pratica di un esperimento. Il prossimo capitolo affronterà questa tematica, individuando una

possibile risposta negli *indici di dimensione dell'effetto*.

Tabella 2.1

Dati ipotetici tra il dosaggio di vitamina C e i giorni passati con il raffreddore (ripresa da Loftus, 2002)

Dati (n = 10.000/ condizione sperimentale)	
Dosaggio quotidiano di vitamina C	Media giorni con il raffreddore
2 g	9.79
3 g	9.72
4 g	9.56

CAPITOLO 3

Gli indici di dimensione dell'effetto

L'ultimo tema affrontato da questo lavoro riguarda l'analisi della *grandezza dell'effetto,* dall'inglese *Effect Size* (*ES*). Il capitolo inizierà dando una definizione al costrutto, proseguirà con una classificazione dei vari tipi di stime riguardo questa dimensione e si concluderà con una lista di alcune raccomandazioni per una pubblicazione efficace.

3.1 Scopo e definizione dell'Effect Size

Nei capitoli precedenti sono state esposte diverse debolezze dell'approccio di verifica dell'ipotesi nulla, da ciò è stato possibile evidenziare come la sola significatività statistica non sia un criterio sufficiente per poter interpretare correttamente gli esiti di una ricerca.

I suggerimenti sull'utilizzo di stime riguardanti la dimensione dell'effetto non possono essere definite "recenti", infatti Gigerenzer (1994) scrive che l'analisi della potenza e gli indici di Effect Size vengono spesso riportati nei libri di testo ma tendono a "scomparire" quando si dovrebbe passare alla pratica. Nel 1996 invece, la Task Force on Statistical Inference (TFSI), chiamata ad esprimersi in merito alle controversie dell'NHST, ha "incoraggiato" i ricercatori a riportare nei propri lavori le stime sull'ampiezza degli effetti studiati. Malgrado ciò, come riporta Thompson (1998), questi incoraggiamenti non sono stati presi molto in considerazione; per questa ragione nel rapporto del 1999 fatto da Wilkinson quelli che prima erano solo dei semplici suggerimenti, diventarono vere e proprie necessità per lo svolgimento e la pubblicazione di un corretto studio scientifico (Wilkinson e la TFSI, 1999). Raccomandazioni "vecchie", che rimangono attuali a causa della, ancora oggi troppo diffusa, noncuranza nei confronti di queste stime.

Tutto ciò però non dà una risposta a cosa sia la *dimensione dell'effetto*. Nakagawa e Cuthill (2007) riportano tre significati che appaiono in letteratura riguardo questo argomento: 1) quello che gli autori chiamano "effetto statistico",

cioè l'insieme degli indici che stimano questa dimensione, 2) il valore reale calcolato dagli indici ed infine 3) l'interpretazione dei risultati nei termini di importanza pratica.

Ferguson (2009), invece, nel suo articolo "guida" all'uso degli indici di *effect size*, spiega che essi servono per evidenziare la grandezza di un effetto pratico, oppure la misura dell'associazione tra due variabili, attraverso delle stime. In questa definizione l'accento va messo alla parola "stime", poiché così come per gli altri risultati ottenuti attraverso le statistiche, esse vanno interpretate nelle situazioni specifiche. Non essendo influenzate dalla numerosità campionaria, a differenza del *p-value* (vedi paragrafo 1.2.3), forniscono delle misurazioni considerabili come veritiere riguardo il fenomeno studiato (Ferguson, 2009). Dopotutto il *p* e gli indici di *ES* rispondono a quesiti molto diversi, il primo comunica soltanto se un fenomeno è presente o meno, i secondi invece in che misura il fenomeno è presente.

3.2 Le famiglie di indici

Attualmente esistono moltissimi indici di dimensione dell'effetto, e persino diverse tipologie di classificazioni; riprendendo la divisione fatta da Kraemer et al. (2003) possiamo definire due gruppi:

- *d family,* gli indici appartenenti a questa categoria vengono utilizzate quando si confrontano le medie di due gruppi e, se pur con formule diverse, esprimono tutti la differenza tra le medie in unità di deviazioni standard. Questa famiglia di stime prende il nome dal suo indice più noto, la *d* di Cohen.
- *r family*, questa famiglia di indici, invece, viene usata per misurare la forza di associazione tra due variabili, tra queste è presente la *r* di Pearson. Questa famiglia possiede un range teorico che oscilla tra -1 e +1, dove lo 0 rappresenta l'assenza di un qualsiasi effetto.

In queste "famiglie" non rientra tutta la gamma di indici possibili: per i fini di questo lavoro verrà approfondita solo la prima di quelle sopracitate, ciononostante, in vista delle debolezze evidenziate nel capitolo precedente, anche una classificazione

semplicistica come questa può essere utile per mostrare che non esiste una misura che sia universalmente affidabile e applicabile in ogni situazione. Compreso ciò, sorge una delle questioni più opprimenti per i ricercatori: qual'è il limite per considerare "significativo" un indice di dimensione dell'effetto?

Come per il *p-value*, anche per gli indici di ES non esiste un limite che sia ontologicamente valido in qualsiasi circostanza, tuttavia esistono alcuni criteri convenzionali che possono aiutare un ricercatore alle prime armi come linee guida nell'interpretazione dei risultati. Nella scelta di queste convenzioni uno degli autori maggiormente considerato è senz'altro Cohen (1992b), che nel suo articolo propone una serie di valori generalmente indicanti un effetto di piccole, medie o grandi dimensioni. La tabella 3.1 contiene i valori suggeriti per due degli indici più noti, la *d* di Cohen e la *r* di Pearson.

Tabella 3.1

Valori convenzionali per l'interpretazione degli indici di ES (riadattata da Cohen, 1992b)

Indice di ES	Effetto piccolo	Effetto medio	Effetto grande
d	.20	.50	.80
r	.10	.30	.50

3.2.1 Differenze medie standardizzate

Ai fini di questo lavoro una dettagliata spiegazione di ognuna delle "famiglie" sopracitate sarebbe eccessivamente lunga, per questo motivo verrà approfondita quanto più esaustivamente possibile solo la prima categoria, in modo da delineare un quadro che, se pur generico, permetta di comprendere l'importanza che questo insieme di indici statistici ha nella ricerca scientifica attuale.

La *d family* riguarda le "differenze medie standardizzate" ed il motivo per cui è necessaria una standardizzazione delle misure è alquanto ovvio: in psicologia non esiste una misurazione universalmente accettata per le dimensioni astratte che indaga, persino uno stesso costrutto spesso possiede diverse tipologie di misurazioni. Il processo di standardizzazione permette di confrontare risultati con unità di misura

differenti trasformandoli in una unità di misura comune. Tutti gli indici di questa famiglia vengono espressi in unità di deviazioni standard. In tal modo, un eventuale rilevamento di un effetto di .20, significherebbe che il gruppo sperimentale ha un andamento con una deviazione standard più grande di un quinto rispetto al gruppo di controllo (Vacha-Haase e Thompson, 2004). La formula generale che caratterizza questi indici risulta:

$$(M_1 - M_2) / \sigma^*$$

Dove M_1 ed M_2 sono le medie del gruppo di controllo e del gruppo sperimentale, mentre σ^* corrisponde ad una stima della deviazione standard della popolazione (questa stima cambia a seconda dell'indice che viene scelto).

3.3 Raccomandazioni finali

Concludendo, questo lavoro intende seguire l'esempio offerto da alcuni autori nel fornire poche, se pur efficaci, raccomandazioni per un buon utilizzo ed una buona pubblicazione delle ricerche scientifiche.

La prima, chiaramente, non può prescindere da quanto scritto nel rapporto fatto da Wilkinson e la TFSI (1999), e cioè: *quando si intende riportare un p-value bisogna sempre accompagnarlo con delle stime sull'ampiezza dell'effetto*. Una norma entrata a far parte anche del Manuale di Pubblicazione dell'American Psychological Association.

La seconda raccomandazione, per quanto ovvia possa sembrare, è: *esplicitare esattamente quale indice si intende usare* (Vacha-Haase e Thompson, 2004). Nei paragrafi precedenti è stato detto come le modalità di interpretazione varino a seconda sia dei diversi indici sia dei diversi contesti che utilizzano lo stesso indice; per questa ragione sono facilmente intuibili le complicazioni che nascerebbero nel caso in cui venisse pubblicata esclusivamente la dicitura “effect size = .50”. La tipologia di comunicazione assume significati profondamente diversi anche considerando solamente i valori convenzionali di *d* o di *r*.

Tener presente che le stime di ES possono essere influenzate da diversi problemi metodologici (Ferguson, 2009), ad esempio le modalità di campionamento o di raccolta dei dati, se il campione oggetto di indagine fosse molto piccolo o se

questo non fosse randomizzato. Il ricercatore, nella rilevazione di un effetto, dovrebbe essere cauto al momento di interpretare i risultati, evitando così delle generalizzazioni erronee.

L'ultima raccomandazione è forse la più complicata e riguarda la necessaria *interpretazione degli indici di ES (così come dei valori del p-value)*. La difficoltà di quest'ultimo punto risiede nella non univocità delle interpretazioni riguardo le stime dell'effetto. Il loro significato pratico è, infatti, strettamente legato al contesto in cui vengono utilizzate e perciò richiedono un presa di posizione da parte del ricercatore. Affinché vengano trasmesse ai lettori le implicazioni che uno studio può avere è fondamentale non limitarsi a riportare i valori di Effect Size, ma integrarle con un commento ragionato su quanto ottenuto.

CONCLUSIONI

L'obiettivo di questa tesi consiste nel riportare alcune nozioni riguardanti l'inferenza statistica nelle ricerche condotte nel campo delle scienze umane, in particolare della psicologia. Attraverso una revisione della letteratura è stato possibile focalizzare l'attenzione su due particolari tematiche: il senso della significatività statistica e gli indici di dimensione dell'effetto.

Per prima cosa è stata descritta la procedura di indagine statistica che attualmente viene impiegata nelle sperimentazioni in psicologia: la *verifica dell'ipotesi nulla* (NHST). L'analisi inferenziale condotta con questa metodologia consiste nella ricerca della significatività statistica in un confronto tra due ipotesi.

In passato, per quanto non sia scomparsa del tutto, era comune la pratica di pubblicare una ricerca esclusivamente se riportava dei risultati statisticamente significativi. Grazie ad una panoramica delle criticità che accompagnano l'NHST è stato possibile evidenziare che la significatività statistica permette di stabilire *se* un fenomeno è presente, ma non fornisce alcuna informazione su *quanto ampio* sia, ne se esso abbia una qualche importanza pratica. Per ottenere informazioni riguardo questi aspetti è necessario ricorrere alle stime di *dimensione dell'effetto* (*Effect Size*).

Le diverse tipologie di indici di Effect Size consentono di valutare l'ampiezza di un effetto in ogni disegno di ricerca; la loro difficoltà risiede nella scelta dell'indice più adatto e nell'interpretazione dei risultati.

Purtroppo non è stato possibile condurre una trattazione dettagliata riguardo queste stime, ciononostante la presente tesi si augura di incoraggiare i ricercatori, presenti e futuri, all'uso consapevole ed integrato dei diversi strumenti statistici che sono a disposizione.

RINGRAZIAMENTI

Vorrei ringraziare prima di tutto la mia famiglia, senza la quale non sarei chi sono oggi. A mio padre e mia madre, per avermi sempre sostenuto e per la loro capacità di farmi sentire al sicuro solamente con la loro approvazione. A mio fratello, per esserci sempre stato. Al resto dei miei parenti: zii, zie, nonni e cugini, per avermi regalato una infanzia piena di risate.

Un caloroso ringraziamento va anche a tutti i miei amici, per essere degli splendidi compagni di avventura. Agli ex compagni del liceo, per la gioia che riescono a regalare ad ogni rimpatriata. Agli amici nuoresi in trasferta universitaria, per essere sempre stati in grado di sopportarmi. Ai colleghi, in particolar modo a Monica, Mattus e Joe, che mi hanno accompagnato in questo percorso formativo aiutandomi a crescere come pochi altri hanno fatto.

Infine, vorrei dedicare un pensiero anche al mio relatore, professor Altoè, e al professor Nicotra, per i buoni consigli e per la disponibilità dimostrata nei miei confronti.

BIBLIOGRAFIA

Abelson, R. P. (1997). On the surprising longevity of flogged horses: Why there is a case for the significance test. *Psychological Science, 8*(1), 12-15. doi: 10.1111/j.1467-9280.1997.tb00536.x

Agnoli, F., & Furlan, S. (2008). La differenza che fa la differenza: dalla significatività statistica alla significatività pratica. *Psicologia clinica dello sviluppo, 12*(2), 211-246. doi: 10.1449/27506

Agnoli, F., & Furlan, S. (2009). I cambiamenti nella verifica di ipotesi: statistiche migliori per decisioni migliori. *Giornale italiano di psicologia, 36*(4), 849-882. doi: 10.1421/30940

Bakan, D. (1966). The test of significance in psychological research. *Psychological bulletin, 66*(6), 423-437. doi: 10.1037/h0020412

Bernstein, S., Bernstein, R. (1999) *Elements of Statistics II – Inferential Statistics*. The McGraw-Hill Companies, Inc. (trad. it. *Statistica Inferenziale*, The McGraw-Hill Companies, srl, Milano, 2003)

Carver, R. P. (1978). The case against statistical significance testing. *Harvard Educational Review*, *48*(3), 378-399.
Recuperato da: http://her.hepg.org/home/main.mpx

Christensen, R. (2005). Testing Fisher, Neyman, Pearson, and Bayes. *The American Statistician*, *59*(2), 121-126. doi: 10.1198/000313005X20871

Cohen, J. (1992a). Statistical power analysis. *Current directions in psychological science*, *1*(3), 98-101. doi: 10.1111/1467-8721.ep10768783

Cohen, J. (1992b). A power primer. *Psychological bulletin*, *112*(1), 155-159. doi: 10.1037/0033-2909.112.1.155

Cohen, J. (1994). The earth is round (p<. 05). *American psychologist*, *49*(12), 997-1003. doi: 10.1037/0003-066X.49.12.997

Daniel, L. G. (1998). Statistical significance testing: A historical overview of misuse and misinterpretation with implications for the editorial policies of educational journals. *Research in the Schools*, *5*(2), 23-32.
doi: 10.1.1.122.4807

Frick, R. W. (1996). The appropriate use of null hypothesis testing. *Psychological*

Methods, *1*(4), 379-390. doi: 10.1037/1082-989X.1.4.379

Gigerenzer, G. (1998). We need statistical thinking, not statistical rituals. *Behavioral and Brain Sciences*, *21*(2), 199-200 doi: 10.1017/S0140525X98281167

Gigerenzer, G. (2004). Mindless statistics. *The Journal of Socio-Economics*, *33*(5), 587-606. doi: 10.1016/j.socec.2004.09.033

Gigerenzer, Krauss, Vitouch (2004). The Null Ritual. What You Always Wanted to Know About. Significance Testing but Were Afraid to Ask. *Published in: D. Kaplan (Ed.). (2004). The Sage handbook of quantitative methodology for the social sciences (pp.391–408).* Recuperato da: http://books.google.it/books

Haller, H., & Krauss, S. (2002). Misinterpretations of significance: A problem students share with their teachers. *Methods of Psychological Research*, *7*(1), 1-20. Recuperato da: http://www.mpr-online.

Hubbard, R., & Bayarri, M. J. (2003). P values are not error probabilities. *Institute of Statistics and Decision Sciences, Working Paper*, (03-26), 27708-0251. Recuperato da: http://ftp.stat.duke.edu/WorkingPapers/03-26.html

Huberty, C. J. (1993). Historical origins of statistical testing practices: The treatment of Fisher versus Neyman-Pearson views in textbooks. *The Journal of Experimental Educational, 61(4)*, 317-333. Recuperato da: http://www.jstor.org/stable/20152388

Huck, S. W. (2012). *Reading statistics and research (6th Ed.)*. Pearson.

Kirk, R. E. (2001). Promoting good statistical practices: Some suggestions. *Educational and Psychological Measurement*, *61*(2), 213-218. doi: 10.1177/00131640121971185

Kraemer, H. C., Morgan, G. A., Leech, N. L., Gliner, J. A., Vaske, J. J., & Harmon, R. J. (2003). Measures of clinical significance. *Journal of the American Academy of Child and Adolescent Psychiatry*, *42*(12), 1524. doi: 10.1097/00004583-200312000-00022

Loftus, G. R. (2002). Analysis, interpretation, and visual presentation of experimental data. *Stevens' handbook of experimental psychology, 4*. doi: 10.1002/0471214426

Luccio, R., Salvadori, E., & Bachmann, C. (2005). *La verifica della significatività dell'ipotesi nulla in psicologia*. Firenze: Firenze University Press. Recuperato

da: *http://digital.casalini.it/8884532264*

Nakagawa, S., & Cuthill, I. C. (2007). Effect size, confidence interval and statistical significance: a practical guide for biologists. *Biological Reviews*, *82*(4), 591-605. doi: 10.1111/j.1469-185X.2007.00027.x

Nickerson, R. S. (2000). Null hypothesis significance testing: a review of an old and continuing controversy. *Psychological methods*, *5*(2), 241-301. doi: 10.1037/1082-989X.5.2.241

Pastore, M. (2009). I limiti dell'approccio NHST e l'alternativa Bayesiana. *Giornale italiano di psicologia*, *36*(4), 925-940. doi: 10.1421/30944

Rosnow, R. L., & Rosenthal, R. (1989). Statistical procedures and the justification of knowledge in psychological science. *American Psychologist*, *44*(10), 1276-1284. doi: 10.1037/0003-066X.44.10.1276

Schmidt, F. L. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers. *Psychological methods*, *1*(2), 115-129. doi: 10.1037/1082-989X.1.2.115

Thompson, B. (1998). Statistical significance and effect size reporting: Portrait of a possible future. *Research in the Schools*, *5*(2), 33-38. Recuperato da: http://www.personal.psu.edu/users/d/m/dmr/sigtest/master.pdf#page=41

Thompson, B. (2002). "Statistical, ""practical, " and "clinical": How many kinds of significance do counselors need to consider?. *Journal of Counseling & Development*, *80*(1), 64-71. doi: 10.1002/j.1556-6678.2002.tb00167.x

Vacha-Haase, T., & Thompson, B. (2004). How to estimate and interpret various effect sizes. *Journal of counseling psychology*, *51*(4), 473-481. doi: 10.1037/0022-0167.51.4.473

Vellone, E., & Alvaro, R. (a cura di). (2011). *Manuale di pubblicazione dell'American Psychological Association*. Napoli, Italia: Edises.

Wagenmakers, E. J. (2007). A practical solution to the pervasive problems ofp values. *Psychonomic Bulletin & Review*, *14*(5), 779-804. doi: 10.3758/BF03194105

Welkowitz, J., Cohen, B., & Ewen, R. (2006). *Introductory Statistics for the Behavioral Sciences*. John Wiley & Sons, Inc. (trad. it. *Statistica per le scienze del comportamento*, APOGEO, srl, Milano, 2009)

Wilkinson, L., and the Task Force on Statistical Inference (1999). Statistical methods in psychology journals: Guidelines and explanations. *American psychologist*, *54*, 594-604. doi: 10.1037/0003-066X.54.8.594