# 1

# Il problema della credibilità dei risultati in psicologia: quale il contributo della statistica?

Francesca Lionetti[(1)], Gianmarco Altoé [(2)], Massimiliano Pastore[(2)]

[(1)] Queen Mary University of London, [(2)] Università di Padova

**Sommario** Nell'ultimo decennio la ricerca in psicologia è stata accusata di risultati scarsamente affidabili e non replicabili. I responsabili di questa crisi di credibilità vengono identificati in pratiche di ricerca eticamente discutibili o in un approccio ai dati in buona fede ma basato su una conoscenza non aggiornata dei metodi statistici nonostante il crescere in complessità dei modelli teorici analizzati. Attraverso una revisione della letteratura e due esemplificazioni, l'articolo introduce il dibattito sulla credibilità dei risultati e discute del valore che buone prassi di analisi dati possono dare al progredire delle conoscenze in un reciproco arricchirsi di saperi tra psicologia e statistica.

**Parole chiave**: crisi di credibilità, analisi dei dati, psicologia, statistica

*Psychology's credibility crisis. Does statistic matter?*

In the last ten years psychology has been accused of non-reliable and non-replicable results. The scientifical debate had acknowledged Questionable Research Practices and a non-updated knowledge of statistic techniques, in front of a growing complexity of research models and data, as the main responsible of this credibility crisis. Through a literature review and practical exemplifications, the current paper introduces the debate on psychology's credibility crisis, and discusses the contribution that reliable and updated techniques of data analysis may give in response to this credibility crisis.

**Key-words**: credibility crisis, data analysis, psychology, statistics

Printed: 04-05-2017

Please cite as: Lionetti, F., Altoè, G., Pastore, M. (2017) *Psicologia Clinica dello Sviluppo*, 21, 3-14.

## Premessa

Nell'ultimo decennio alcune discipline che si basano sull'analisi dei dati per il progredire delle conoscenze, o *statistics related fields*, sono state accusate di scarsa credibilità per la presenza negli articoli di risultati spesso poco chiari, parziali, non generalizzabili al di fuori dal contesto del laboratorio e la cui conferma viene meno al momento della replica (Ioannidis, 2005). Tra le discipline coinvolte in questa crisi, la psicologia si discosta particolarmente da altri settori della ricerca per l'elevata percentuale di risultati statisticamente significativi pubblicati, aspetto che, sebbene non escluda un possibile bias da parte delle riviste nelle scelte editoriali o una tendenza dei ricercatori a inviare per il processo di revisione prevalentemente articoli che confermano le proprie ipotesi, ha tuttavia generato non poche osservazioni critiche circa l'affidabilità dei risultati (Fanelli, 2012). Trascorsi dieci anni dall'inizio dell'attenzione scientifica e mediatica su questo tema, e in concomitanza a un recente lavoro su replicabilità e credibilità pubblicato sulla celebre rivista Science (Open Science Collaboration, 2015) che suggerisce come il dibattito sia tutt'altro che sopito, il presente lavoro si propone di contribuire a quel movimento di parte *construens* che ha iniziato a prendere piede come reazione alla crisi di credibilità, suggerendo che un approccio metodologico e statistico *data-driven*, ossia guidato dai dati osservati, possa essere una tra le risposte a questa crisi. La prima parte dell'articolo è dedicata a una sintesi dei contributi più rilevanti sul tema, e si conclude con una proposta di linee guida derivate dalla letteratura sicentifica tra prassi consolidate e nuovi spunti di riflessione. La seconda parte è invece di tipo applicativo, e presenta due esemplificazioni di analisi in contesti tipici della psicologia, con l'obiettivo di mostrare nella pratica come alcune specifiche tecniche possano contribuire ad ottenere, in modo più coerente, maggiori informazioni dai dati.

## Statistics related fields e crisi di credibilità

L'accusa che più di altre mina la credibilità della ricerca è l'impiego di *Questionable Research Pratice* o, come li definiscono John, Loewenstein, e Prelec (2012), steroidi della competizione accademica; ad esempio includere nell'articolo solo i risultati statisticamente significativi, interrompere il reclutamento soggetti non appena l'ipotesi venga confermata, o escludere casi per aumentare la probabilità che il risultato vada nella direzione attesa. In sintesi, pratiche volte a diffondere sul mercato della scienza un risultato che sia citato, di successo, e con maggior probabilità accettato dalle riviste, obiettivo legittimo ma perseguito in modo non necessariamente etico. Si tratta di critiche diffuse in ambito accademico ma che hanno raggiunto anche l'attenzione dei media: ha fatto scalpore il caso riportato il 18 agosto del 2015 sul quotidiano americano Washington Post in cui veniva citata una famosa casa editrice, cui appartengono diverse riviste indicizzate anche della psicologia, che ha ritirato 64 articoli per processi di revisione falsati, ossia effettuati da *fake account*, identità false (o identità di esperti vere nel nome ma associate a credenziali false) inserite come opzioni tra i revisori suggeriti dagli autori durante l'invio di un lavoro a una rivista tramite le note piattaforme informatiche. Pratiche non condivisibili sul piano etico, e che meritano riflessione rispetto alla direzione verso cui la psicologia - o in generale la ricerca - si sta muovendo, sono sicuramente ciò che più desta attenzione e genera dibattito.

Accanto a questa mala prassi vi sono processi inconsapevoli di approccio ai dati che altrettanto minacciano la credibilità dei risultati; diversi certamente come valenza sul piano etico sono in egual misura oggetto di critica perché in modo pur non volontario rischiano di minare la credibilità delle conclusioni tratte dagli studi. Ci riferiamo in particolare a una cultura metodologica che si protrae in buona fede ma in modo poco aggiornato e con scarso confronto interdisciplinare (Maxwell, 2004; Ioannidis, 2005; Young, Ioannidis, & Al-Ubaydi, 2008; Bakker & Wicherts, 2011; Simmons, Nelson, & Simonsohn, 2011; Cumming, 2012; Masicampo & Lalande, 2012; Schimmack, 2012; Francis, 2013; ?, ?; Gigerenzer & Marewski, 2015; Open Science Collaboration, 2015), portando al frequente utilizzo di statistiche del XX secolo nel XXI (Gelman, 2015) che, con una metafora, potremmo paragonare all'utilizzo di una automobile degli anni '30 del secolo scorso nel traffico attuale (Pastore, 2014). Crescono le conoscenze teoriche, vi è un proliferare di contributi che propongono nuovi e più articolati modelli di

comprensione del funzionamento dei processi cognitivi, emotivi, sociali, crescono in complessità i modelli in esame ma spesso non lo stesso aggiornamento si riscontra nell'analisi dei dati (Sijtsma, 2009); così un modello di spiegazione del funzionamento della mente viene abbandonato quando la ricerca lo disconferma o ne individua uno migliore, ma non lo stesso accade con i modelli statistici applicati in psicologia.

La risposta costruttiva a questa crisi non ha mancato di farsi sentire e un vivace dibattito si è avviato anche in Italia, con particolare attenzione al ruolo degli studi di replicabilità e all'utilizzo di tecniche di analisi più opportune e rispettose del dato (Della Sala & Cubelli, 2014; Pastore, 2014; Perugini, 2014). Sia a livello nazionale, sia a livello internazionale, la cultura metodologico-statistica è stata in prima linea attiva e partecipe in questo dibattito. Tra le proposte di analisi dei dati diffusesi in questo contesto un posto d'onore è sicuramente occupato dall'approccio bayesiano (es. Kruschke, 2011) come via per limitare la tendenza alla ricerca della significatività statistica ad ogni costo, aspetto tra i più criticati in questo decennio (Ioannidis, 2005; Ziliak & McCloskey, 2008; Open Science Collaboration, 2015). Accanto all'approccio bayesiano, altri elementi sicuramente estendibili anche al più classico approccio frequentista quali l'attenzione posta al trattamento dei dati mancanti - comuni nei disegni longitudinali (Peeters, Zondervan-Zwijnenburg, Vink, & van de Schoot, 2015) -, la riflessione sulla migliore numerosità campionaria in base al disegno di ricerca oltre la semplice analisi della potenza, nonché l'interesse per come aumentare la scientificità dei disegni di ricerca attraverso la manipolazione delle variabili indipendenti anche negli studi della psicologia dello sviluppo (Bakermans-Kranenburg & van IJzendoorn, 2015), testimoniano della crescente attenzione per il rigore metodologico che la psicologia ha mostrato in risposta a questa crisi. È interessante citare più nel dettaglio, in questo processo di riflessione costruttiva, il recente lavoro apparso sulla rivista Science introdotto nella premessa all'articolo: un gruppo di 100 studiosi ha cercato di replicare i risultati di 270 studi pubblicati su riviste di alto impatto scientifico, e solo il 36% ha portato allo stesso risultato (Open Science Collaboration, 2015), a supporto di quanto Ioannidis nel 2005 aveva ipotizzato. Ciò che è rivoluzionario in questo lavoro, che appare solo a prima vista un ennesimo indicatore di crisi, è il dibattito che ne è derivato e l'apprezzamento per il confronto, rapidamente diffusosi nei forum scientifici, proprio da parte degli autori dei lavori originali. Dopo un decennio di crisi (Ioannidis, 2005; Open Science Collaboration, 2015) la psicologia si è mostrata pronta per discutere apertamente delle sue ombre, motivata alla necessità della replica e del confronto, cauta nel generalizzare le conclusioni da un solo studio, e umile nel confronto scientifico che permette la crescita, confermando l'avvio di un processo sempre più bi-direzionale di scambio tra studio dei processi psicologici e rigore metodologico.

### Proposte costruttive, nuovi miti ed esiti paradossali

L'approccio bayesiano, che sopra abbiamo citato come una tra le ultime piccole rivoluzioni di pensiero importate nell'analisi dei dati in psicologia, ha avuto il merito di favorire il dibattito sulla correttezza delle conoscenze metodologiche e statistiche più diffuse nelle scienze psicologiche e sociali (Altoé & Pastore, 2013; Klugkist, van Wesel, & Bullens, 2011; Kruschke & Liddell, 2017). Vorremmo tuttavia richiamare l'attenzione su come ogni soluzione, se assunta come nuovo dogma, corra gli stessi rischi dell'uso acritico del valore soglia di $p$. Ne è esempio l'editoriale della rivista *Basic Applied Social Psychology* (Trafimow & Marks, 2015) in cui viene bandito l'uso del *p-value* arrivando al paradosso che, qualora accettati, gli articoli che hanno basato le loro analisi sull'approccio frequentista classico eliminino il riferimento alla significatività pur avendo su questa basato le loro conclusioni, ossia si arriva al paradosso di utilizzarlo senza ammetterlo, e di accettare un contributo esito di quella modalità di approccio ai dati richiedendo tuttavia infine di negarla. In reazione a questo editoriale numerosi autorevoli ricercatori, che pure hanno avuto parole di forte criticità verso l'approccio NHST (*Null Hypothesis Significance Testing*; Cohen, 1994), sono intervenuti sottolineando come il rischio di bandire in toto dalle pubblicazioni l'approccio frequentista sia di introdurre una nuova prassi dogmatica come dato di fatto e nuova fede metodologica (Flanagan, 2015).

Ancora prima della diffusione del metodo bayesiano in psicologia, lo stesso destino è toccato alla grandezza dell'effetto ed agli intervalli di confidenza. Tale è stata la diffusione di questi indicatori,

intuitivi ed immediati, baluardi di quella che è stata definita come *new statistic* (Cumming, 2012), che si è giunti a esiti paradossali di passaggio da un dogma (quello del *p-value*) a un altro: Gigerenzer e Marewski (2015) citano un articolo in cui persino la numerosità campionaria dei soggetti veniva riportata con un proprio intervallo di confidenza. Aspetto ancora più paradossale è che la stima degli intervalli di confidenza, proposti come alternativa per ovviare ai limiti di NHST, si basa sullo stesso calcolo da cui si ottiene il *p-value* e, sebbene il loro utilizzo sia fortemente suggerito dalle norme APA (American Psychological Association) sin dal 2001, esso risulta ancora soggetto a fallaci interpretazioni, come dimostrato in una recente ricerca che ha coinvolto oltre 100 ricercatori e 400 studenti di psicologia (Hoekstra, Morey, Rouder, & Wagenmakers, 2014). Se consideriamo invece la pratica di riportare la grandezza dell'effetto, si osservano ancora raramente autori che commentano il significato del suo valore e spesso questo accade anche quando contraddice con la sua grandezza le conclusioni che il valore del $p$ suggerisce (Cumming et al., 2007; Fidler et al., 2005; Finch et al., 2004). Il problema dunque non sembra essere esclusivamente l'approccio che si utilizza in sè, quanto un'inferenzialità nel trarre conclusioni dai dati che corre il rischio di essere semplicistica, acritica e poco ragionata, sia essa applicata in ottica frequentista, bayesiana o altro. Ecco in cosa consiste dunque la cassetta degli attrezzi che di seguito proponiamo: una serie di riflessioni di cui tener conto prima di applicare in modo automatico una tecnica di analisi dei dati, affinché siano il rigore metodologico, il dato, e l'ipotesi, a guidare le analisi piuttosto che una statistica dogma assunta a priori.

## La cassetta degli attrezzi

Quali strumenti può essere in particolare utile inserire nella cassetta degli attrezzi per condurre valide e affidabili analisi? Cosa è utile contenga la nostra *statistical toolbox* - citando Gigerenzer e Marewski (2015) - in quanto ricercatori in psicologia? Proviamo ad individuare questi elementi; alcuni sono certamente strumenti di lavoro tradizionali, come si evince anche dal riferimento ai classici della letteratura, e che derivano dalla definizione stessa di modello statistico come sistema formale di assunti che deve essere rispettato affinché i risultati ottenuti abbiano senso e siano interpretabili; altri fanno riferimento ad aspetti metodologici della ricerca, come ad esempio l'importanza di riportare tutti i dati, anche quelli non statisticamente significativi. Vediamoli ora più nel dettaglio.

Un primo strumento utile, e noto da tempo, è l'esplorazione grafica dei risultati: la sintesi numerica nell'output delle analisi non sempre è rappresentativa della distribuzione reale dei dati rispetto alla variabile indagata. Come sarà familiare ai più, uno stesso valore della statistica di correlazione $r$ di Pearson può associarsi a distribuzioni dei dati molto differenti o addirittura in netta contrapposizione tra loro; si veda il classico esempio di Anscombe (1973), riportato in figura 1.1, in cui allo stesso valore di correlazione sono associate relazioni tra variabili decisamente diverse tra di loro. La rappresentazione grafica ha un ruolo fondamentale nel processo di analisi statistica e risulta utile a prescindere dall'approccio che si intende utilizzare (Tukey, 1977; Lindley, 2000).

Altra raccomandazione da diversi anni ribadita in letteratura, e parte del patrimonio conoscitivo della ricerca, è l'importanza del porre attenzione alla natura dei dati, in particolare al tipo di scala di misura. In psicologia molte variabili considerate come dipendenti sono ordinali, ma vengono trattate nell'analisi dei dati come continue. Questo può portare a perdere informazione dai dati, penalizzando il ricercatore nelle conclusioni che dal suo studio può trarre, sopratutto quando queste non hanno una distribuzione simmetrica (Jamieson, 2004; Flora & Curran, 2004; Gadermann, Guhn, & Zumbo, 2012).

Per quanto riguarda poi il metodo bayesiano, ampiamente già citato, le assunzioni a priori (o *prior*) su cui si basano le analisi dovrebbero essere riportate per rendere trasparente il risultato finale, così come esplicitati i confronti tra modelli in ottica semi-bayesiana (Gelman, Carlin, Stern, & Rubin, 2014). Qualora il ricercatore assuma un approccio di analisi di tipo frequentista, non dovrebbe essere la regola quella di basarsi su un unico livello di significatività o *p-value* per rigettare l'ipotesi nulla, ma valutare rispetto al contesto il grado di significatività più adeguato, come già sostenuto dallo stesso Fisher (1956). Laddove poi la numerosità campionaria in relazione al numero di parametri stimati lo permetta, considerare modelli unici che tengano in considerazione l'influenza reciproca delle variabili consente di avere un quadro più aderente alla realtà dei processi che spiegano il fenomeno indagato

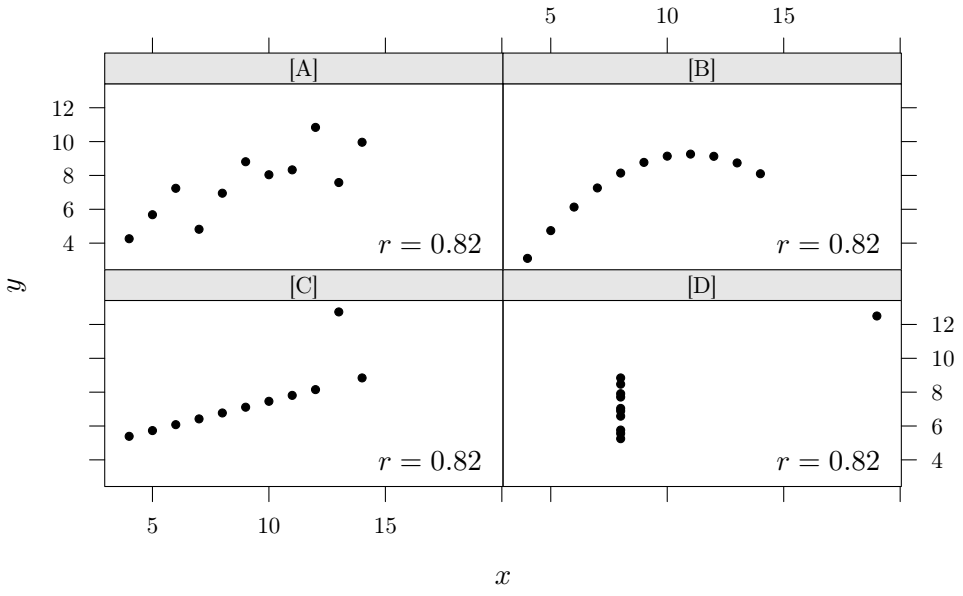


Figura 1.1: Casi esemplari in cui configurazioni di punti molto diverse producono la stessa correlazione $r = 0.82$ (Anscombe, 1973).

piuttosto che condurre molte analisi singole non in relazione tra loro (Gelman & Hill, 2006). Infine, presentare tutti i risultati, anche quelli che contraddicono le nostre ipotesi o che, utilizzando una terminologia più diffusa, sono statisticamente non significativi, certamente contribuirebbe a diminiure il sospetto che la psicologia tenda a pubblicare solo ciò che conferma se stessa e le ipotesi formulate (Fanelli, 2012; Pastore, Nucci, & Bobbio, 2015).

## Applicazioni esemplicative

Passiamo ora alla pratica, e nello specifico alla proposta di due esemplificazioni di tecniche di analisi dei dati, ancora non ampiamente diffuse nei vari ambiti di ricerca della psicologia, che possono permettere di ottenere informazioni più esaustive e complete dei fenomeni in esame. Le due applicazioni che presentiamo sono scelte con particolare attenzione verso la psicologia dello sviluppo e la psicologia dell'educazione; le analisi sono condotte con il software R (R Core Team, 2015) e i dati opportunamente simulati a partire da lavori derivati dalla letteratura. Da una breve revisione dei lavori presentati nella sezione Ricerche su Psicologia Clinica dello Sviluppo tra il 2012 e il 2015 abbiamo riscontrato che il 65% dei contributi ha argomentato le proprie conclusioni utilizzando il test $\chi^2$ oppure il test $t$, tecniche diffuse e certamente utili, ma che oggi vedono altrettante utili alternative, che qui proponiamo come potenzialmente più informative. Pertanto abbiamo scelto due esempi in cui illustrare delle alternative a questi test, facilmente riproducibili ad esempio con appositi pacchetti di R (R Core Team, 2015); in particolare presenteremo: 1) una regressione con variabili ordinali e confronto tra modelli (Yee, 2010); 2) un confronto tra gruppi in forma bayesiana (Kruschke, 2015).

### Modelli per lo studio dei fattori di rischio e protezione

Ipotizziamo di voler studiare i fattori di rischio e protezione per l'attaccamento nel contesto dell'adozione. Scegliamo il riferimento teorico dell'attaccamento in quanto familiare a molti, affinché l'esempio

sia intuitivamente comprensibile senza ripercorrere i dettagli della teoria di riferimento. Ipotizziamo nel nostro esempio, ispirato alla letteratura, di mettere a punto un disegno di ricerca in cui rileviamo in un primo tempo, ossia entro i primi mesi dall'adozione, il pattern di attaccamento del caregiver adottivo principale, ed un anno dopo il pattern di attaccamento del bambino verso il caregiver (Lionetti, 2014; Pace, 2014). Raccogliamo inoltre il dato circa l'età del bambino al momento del collocamento adottivo, ossia se entro il primo mese di vita (ipotizzando l'esistenza di un periodo sensibile nello sviluppo) o successivamente. Le variabili ambientali che assumiamo come possibili predittori del pattern di attaccamento del bambino sono dunque due: l'età all'adozione e il pattern di attaccamento del caregiver principale.

Partiamo dall'analisi dell'associazione tra le variabili coinvolte utilizzando delle tabelle di contingenza, ed esploriamo quale associazione si riveli statisticamente significativa. In seguito procediamo con un approccio di *model selection* (Burnham & Anderson, 2004; Burnham, Anderson, & Huyvaert, 2011; Fox, 2008) in cui vengono messi a confronto vari modelli di regressione scegliendo il migliore utilizzando il *Bayesian Information Criterion* (BIC; Schwarz, 1978); si tenga presente che questo tipo di approccio, nonostante l'aggettivazione, non è bayesiano in senso stretto perchè non richiede di definire delle *prior* e non utilizza delle *posterior*. Nei modelli di regressione consideriamo in particolare due predittori all'interno di un unico modello, assumendo come variabili indipendenti età all'adozione e pattern di attaccamento del caregiver primario, e come variabile dipendente l'attaccamento del bambino. Assumiamo che la variabile attaccamento sia ordinale, indicando con 1 (basso livello di rischio / assenza di rischio) il pattern di attaccamento sicuro; con 2 il pattern di attaccamento insicuro e con 3 il pattern di attaccamento disorganizzato/irrisolto, facendo riferimento per questa graduatoria ordinale a quanto ad esempio riportato in letteratura rispetto all'associazione tra insicurezza, disorganizzazione, e presenza di problematiche comportamentali (Fearon, Bakermans-Kranenburg, Van IJzendoorn, Lapsley, & Roisman, 2010).

Supponiamo di basare queste analisi su un campione di 60 partecipanti, ovvero 30 diadi genitore - bambino in adozione, reclutati nel primo anno del collocamento adottivo. Applicando una tradizionale analisi basata sul $\chi^2$ alle relative tabelle di contingenza (vedi tabelle 1.1 (a) e (b)) individuiamo la presenza di un'associazione statisticamente significativa tra attaccamento materno (1 = sicuro, 2 = insicuro, 3 = irrisolto) e fascia di età del bambino all'adozione (entro il primo mese di vita; dal secondo mese di vita ed entro l'anno di età), con la condizione pattern di attaccamento del bambino (1 = sicuro, 2 = insicuro, 3 = disorganizzato). Tuttavia questa prima analisi, prevalentemente descrittiva, presenta una serie di limiti che possiamo così sintetizzare: 1) considerare i nostri predittori separatamente non rappresenta la condizione reale in cui fattori di rischio e protezione interagiscono tra loro nel contribuire a una data condizione di sviluppo; 2) non rappresenta un modello di predittività in termini statistici, ma possiamo solo parlare di associazioni, nonostante il disegno di ricerca che abbiamo appena descritto e le ipotesi stesse implichino una specifica direzione degli effetti; 3) non tiene in considerazione il livello ipotizzato di rischio crescente, ossia la natura ordinale della variabile dipendente, avendo assunto che la disorganizzazione implichi un livello di rischio più elevato dell'insicurezza e ovviamente della sicurezza

(a)

| | età adozione | |
|---|---|---|
| att. bambino | 0 | 1 |
| 1 | 11 | 2 |
| 2 | 3 | 4 |
| 3 | 2 | 8 |

(b)

| | att. materno | | |
|---|---|---|---|
| att. bambino | 1 | 2 | 3 |
| 1 | 11 | 2 | 0 |
| 2 | 1 | 4 | 2 |
| 3 | 1 | 7 | 2 |

Tabella 1.1: (a) Frequenze osservate per l'incrocio pattern attaccamento del bambino (in riga, 1= sicuro, 2= insicuro, 3= disorganizzato) × e età all'adozione (in colonna, 0= entro il primo mese di vita; 1= tra 2 e 12 mesi di vita). $\chi^2_{(2)} = 9.88$, $p = 0.007$; (b) Frequenze osservate per l'incrocio pattern attaccamento del bambino (in riga, 1= sicuro, 2= insicuro, 3= disorganizzato) × e pattern attaccamento materno (in colonna, 1= sicuro; 2= insicuro, 3= irrisolto). $\chi^2_{(4)} = 16.62$, $p = 0.002$.

| | Modello | bic |
|---|---|---|
| m3 | caregiver + età adozione | 35.25 |
| m1 | caregiver | 35.59 |
| m2 | età adozione | 37.63 |
| m4 | caregiver × età adozione | 38.81 |
| m0 | 1 | 42.46 |

Tabella 1.2: Confronto tramite BIC tra i modelli di regressione VGAM nel predire il pattern di attaccamento del bambino.

in una gerarchia del rischio (Lyons-Ruth & Jacobvitz, 2008).

Per ovviare ai limiti sopra elencati procediamo allora applicando una tecnica alternativa, ovvero un'analisi di regressione che assuma una specifica direzione degli effetti, permetta di considerare simultaneamente le variabili del nostro modello e, in particolare, rispetti la natura ordinale della variabile dipendente. Tale modello prende il nome di *Vector Generalized Linear Model* (VGLM: Yee & Wild, 1996; Liu & Agresti, 2005). Si tratta in sintesi di una regressione simile alla logistica, con la differenza che la variabile dipendente si caratterizza per la presenza di più categorie ordinate.

In pratica, data $y$, una variabile dipendente ordinale con $c$ categorie, siamo interessati a modellare

$$P(y = j), j = 1, 2, ..., c$$

ovvero la probabilità che il valore di $y$ cada in una delle $c$ categorie, in funzione di un insieme di predittori $x$ (quantitativi o categoriali). L'obiettivo diventa quindi la stima delle probabilità condizionali piuttosto che delle medie, come avviene nella regressione lineare, ed $y$ viene trattata come vettore $c$-dimensionale.

Una volta specificato il metodo con cui trattare i dati, adottiamo una strategia di *model selection* (Fox, 2008), ovvero definiamo una serie di modelli con la stessa variabile dipendente introducendo per ciascuno di essi gli effetti che vogliamo testare. Pertanto avremo: un modello senza predittori (modello nullo, m0), due modelli con un solo predittore (rispettivamente caregiver, m1, e età all'adozione, m2), e due modelli con entrambi i predittori (senza interazione, m3 e con interazione, m4). Per stabilire quale tra i cinque modelli individuati (cfr. tabella 1.2) sia il migliore nello spiegare i dati osservati utilizziamo il *Bayesian Information Criterion* (BIC; Schwarz, 1978). Questo ci permette di superare i limiti connessi all'approccio frequentista individuando il modello migliore tra quelli presi in esame piuttosto che basare l'interpretazione dei risultati sull'accettare/rifiutare l'ipotesi nulla. In sintesi, il BIC permette di selezionare il modello che supporta meglio i dati tra un insieme di alternative semplicemente scegliendo quello con il valore più basso.

Come possiamo osservare dal confronto tra modelli riportati in tabella 1.2, nonostante l'associazione statisticamente significativa individuata in tabella 1.1(a), l'età all'adozione risulta meno efficace del ruolo del caregiver nello spiegare la probabilità di presentare un dato pattern di attaccamento (37.63 vs. 35.59). In questo modo, ovvero trattando la variabile dipendente come ordinale in un modello di regressione, stiamo stabilendo una corrispondenza più evidente tra le nostre ipotesi (di predittività) e

| età adozione | attaccamento caregiver | 1 | 2 | 3 |
|---|---|---|---|---|
| 0 | 1 | 0.71 | 0.20 | 0.09 |
| 0 | 2 | 0.33 | 0.33 | 0.33 |
| 0 | 3 | 0.35 | 0.33 | 0.32 |
| 1 | 1 | 0.50 | 0.30 | 0.20 |
| 1 | 2 | 0.17 | 0.28 | 0.55 |
| 1 | 3 | 0.18 | 0.29 | 0.53 |

Tabella 1.3: Probabilità stimate di appartenenza ad un gruppo (1 = attaccamento sicuro; 2 = attaccamento insicuro; 3 = attaccamento disorganizzato, in funzione dei predittori del modello m3.

la tecnica di analisi dei dati utilizzata. Sulla base dei valori attesi del modello identificato come migliore (m3, BIC = 35.25, in cui età all'adozione e il pattern di attaccamento del caregiver hanno un effetto addittivo), procediamo infine a stimare le probabilità di appartenenza a una delle tre categorie individuate (sicurezza, insicurezza, disorganizzazione) in base alla combinazione dei predittori. Diversamente dalle precedenti tabelle di contingenza, consideriamo il contributo congiunto delle variabili in gioco, e stante la differenza tra BIC, presente ma contenuta, possiamo approfondire ulteriormente come i predittori contribuiscano al pattern di attaccamento del bambino in modo dettagliato senza ricorrere acriticamente a un unico indicatore. In questo modo evitiamo di basare le conclusioni su di una sola statistica, ed osserviamo il dato. Come riportato in tabella 1.3, possiamo ad esempio osservare che, dato un genitore con attaccamento sicuro si hanno il 70% di probabilità di appartenere al gruppo dei bambini con attaccamento sicuro, mentre la probabilità di avere un attaccamento disorganizzato con un genitore sicuro è solo del 9%. Analogamente, la somma della condizione età all'adozione e attaccamento insicuro o disorganizzato nel genitore è ciò che contribuisce a determinare un attaccamento disorganizzato con oltre il 50% di probabilità.

È importante notare che, data la differenza contenuta in termini di BIC tra il modello che considera come predittore solo il caregiver (m1), e quello che include inoltre l'età all'adozione (m3) (vedi tab. 1.2), avremmo potuto optare anche per il modello m1, considerando solo il ruolo del genitore come determinante. In questo caso è la teoria di riferimento a guidare il ricercatore, purché basi la scelta supportandola con opportuni riferimenti teorici e riporti inoltre i valori BIC associati a tutte le altre opzioni possibili, rendendo in questo modo il risultato più trasparente.

In termini applicativi l'individuazione dei fattori di rischio e protezione additivi potrebbe consentire una migliore focalizzazione nei programmi di intervento delle variabili su cui lavorare per la promozione del benessere e la riduzione del rischio. Inoltre, dal punto di vista della loro generalizzabilità, i modelli VGLM potrebbero rivelarsi più opportuni e più informativi anche in altri casi frequenti nella letteratura psicologica, ogni qual volta cioè ci troviamo a lavorare con dati di natura ordinale, come le scale likert (Jamieson, 2004), o quando la variabile dipendente sia ad esempio l'individuazione di gruppi di soggetti a rischio crescente (Barone, Bramante, Lionetti, & Pastore, 2014).

### Approccio inferenziale bayesiano per la valutazione dei metodi di insegnamento

In un recente e illuminante articolo Gelman (2015) sostiene, tra gli altri, due aspetti fondamentali dell'analisi statistica e dell'interpretazione dei dati: 1) l'effetto di un trattamento può ragionevolmente manifestarsi nel cambiamento della media e/o della varianza della variabile di interesse. I due fenomeni hanno la stessa importanza e devono sempre essere simultaneamente valutati e interpretati; 2) il principale vantaggio dell'approccio bayesiano nell'analisi dei dati, rispetto agli approcci tradizionali (come NHST), è di fornire maggiori informazioni utili su cui basare l'inferenza statistica e l'interpretazione dei dati. Per comprendere come queste due assunzioni ci possano essere utili per trarre il maggior numero di informazioni dalla nostra ricerca, trattiamo ora un caso e un modello di analisi volutamente semplici e agevolmente replicabili. L'estensione a modelli più articolati ha come limite la creatività del ricercatore (Lee & Wagenmakers, 2014) e la conquista di una dimestichezza possibile grazie a un maggior investimento metodologico e statistico a partire dalla formazione.

Supponiamo che un team di ricercatori voglia confrontare l'efficacia di due metodi di insegnamento della psicometria: uno tradizionale e uno innovativo in cui le lezioni teoriche vengono sempre introdotte da un esempio di analisi statistica di un caso reale presentato al computer. A tale scopo, i ricercatori selezionano casualmente due gruppi di studenti omogenei rispetto alle variabili di interesse per lo studio. Al primo gruppo, detto "sperimentale" (SP; $n = 20$), viene fatto seguire il corso con il metodo innovativo, mentre al secondo gruppo, detto di "controllo" (CO; $n = 20$), viene fatto seguire il corso tradizionale. Per rendere maggiormente omogeneo lo studio, entrambi i corsi vengono tenuti dallo stesso docente. Al termine dei corsi viene somministrata ai soggetti la stessa verifica valutata con un punteggio da 0 a 30, in cui valori elevati indicano un'elevata comprensione della psicometria.

Partiamo, come nell'esemplificazione precedente, da un'analisi e interpretazione dei risultati secondo la prospettiva NHST. Ipotizziamo che il ricercatore, per analizzare i dati, abbia seguito questi passi:

utilizzo del t-test per campioni indipendenti per valutare la differenza tra gruppi e del $d$ di Cohen per stimare la grandezza dell'effetto. Nonostante la dimensione dell'effetto in favore del metodo innovativo, $d = 0.42$, possa essere considerata media rispetto ai valori guida suggeriti da Cohen (1988), i risultati non supportano l'esistenza di una differenza statisticamente significativa tra i due metodi ($M_{SP} = 19.2$, $DS_{SP} = 6.54$, $M_{CO} = 16.95$, $DS_{CO} = 3.86$, $t(38) = 1.325$, $p = 0.193$)[1].

Proviamo ora a cambiare prospettiva: rappresentiamo i dati attraverso dei boxplot o "diagrammi a scatola e baffi" (Tukey, 1977) per una prima descrizione dei risultati dello studio (vedi figura 1.2). Guardando il grafico, il primo aspetto che appare evidente è la differenza tra la variabilità dei punteggi nei due gruppi. Nel gruppo sperimentale il minimo e il massimo dei punteggi, rappresentati dagli estremi dei "baffi", sono rispettivamente 8 e 30; mentre nel gruppo di controllo essi sono 10 e 23. Inoltre, osservando il limite inferiore e il limite superiore delle "scatole", rispettivamente il 25-esimo e il 75-esimo percentile della distribuzione dei dati, si può notare che, mentre nel gruppo di controllo il 50% dei dati è all'incirca compreso tra 14 e 20, nel gruppo sperimentale la stessa percentuale di dati è compresa tra 14 e 24. In sostanza, l'effetto più rilevante del trattamento sembra essere quello di aver prodotto una maggiore variabilità nei punteggi nel gruppo sperimentale. Accanto a questo aspetto, i punteggi nel gruppo sperimentale sembrano essere complessivamente più alti rispetto a quelli del gruppo di controllo. In particolare, la mediana (evidenziata dalla linea nera all'interno delle "scatole") risulta circa 19 nel gruppo sperimentale e 17 in quello di controllo.

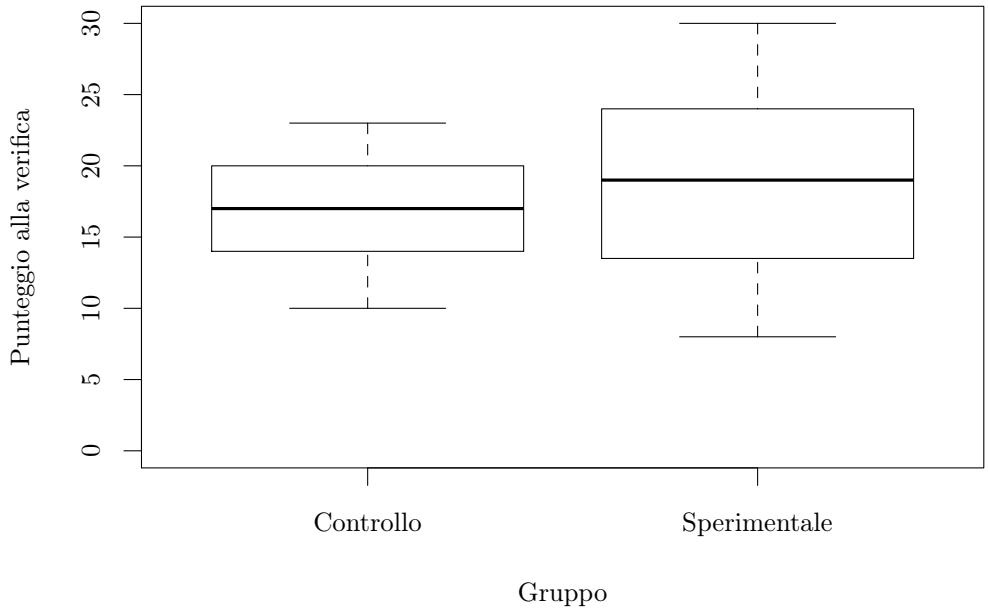


Figura 1.2: Distribuzione dei punteggi al test di verifica ($n = 40$).

Procediamo ora ad analizzare i dati utilizzando un approccio bayesiano (per ulteriori approfondimenti si veda ad es.: Lee & Wagenmakers, 2014; Kruschke, 2015), valutando quali informazioni possiamo trarre oltre questo primo screening a livello descrittivo. In estrema sintesi, nell'approccio bayesiano i parametri oggetto di inferenza (nel nostro caso, le due medie e le due deviazioni standard

[1] Anche nel caso di un'analisi più approfondita, in cui si fosse preventivamente valutata l'omogeneità delle varianze dei due gruppi ($F(1, 38) = 6.389, p = 0.016$), e si fosse proceduto di conseguenza ad utilizzare la versione corretta del t-test ($t(30.794) = 1.325$, $p = 0.195$): 1) la stima del $d$ di Cohen non sarebbe cambiata; 2) le conclusioni non sarebbero cambiate in modo sostanziale.

delle popolazioni da cui provengono i campioni studiati) vengono considerati delle variabili casuali descritte da una distribuzione di probabilità, in contrasto con quanto avviene nell'approccio NHST in cui i parametri sono considerati dei valori fissi. È bene sottolineare che l'assunzione bayesiana permette di valutare le proprietà dei parametri incogniti in termini direttamente probabilistici, garantendo la coerenza dell'analisi (Rouder, Speckman, Sun, Morey, & Iverson, 2009) e aumentando le informazioni ottenibili.

In breve, nell'approccio bayesiano vengono prima definite delle distribuzioni di probabilità a priori, dette *priors*, per i parametri incogniti. Tali prior possono essere vaghe se non si dispone di informazioni a priori, o possono riflettere il grado di conoscenza a priori sul fenomeno di studio e/o le specifiche ipotesi del ricercatore. Nel caso in esame, consideriamo delle prior volutamente vaghe, come se fossimo in assenza di rilevanti informazioni a priori sul fenomeno di studio. È un caso tipico di quelle condizioni in cui non si disponga di studi importanti precedenti che guidino con maggior certezza le attese attuali; scegliendo prior vaghe le stime dei parametri saranno influenzate quasi esclusivamente dai dati osservati. Ipotizziamo che il campione sperimentale provenga da una popolazione normale con media $\mu_y$ e deviazione standard $\sigma_y$, in cui: 1) $\mu_y$ può assumere con la stessa probabilità qualsiasi valore tra 0 (minimo punteggio osservabile) e 30 (massimo punteggio osservabile) ossia, in termini formali, assumiamo che $\mu_y$ sia distribuita come una *variabile casuale uniforme* di parametri 0 e 30; 2) allo stesso modo, $\sigma_y$ ha distribuzione uniforme di parametri 0 e 15 (valutando ragionevole che un limite 'abbondantemente' superiore per la deviazione standard possa essere pari a 15). In modo del tutto analogo costruiamo le prior per i parametri $\mu_x$ e $\sigma_x$ che si riferiscono alla popolazione da cui è tratto il campione di controllo.

È bene sottolineare, fin da ora, che medie e deviazioni standard vengono valutate simultaneamente e che nessuna assunzione viene fatta sull'omogeneità delle varianze: saranno direttamente i dati a suggerirci la fondatezza di tale ipotesi. Utilizziamo quindi un algoritmo di ricampionamento Markov Chain Monte Carlo (MCMC; Geman & Geman, 1984; Gelfand & Smith, 1990; Kruschke, 2015), e osserviamo sulla base dei dati rilevati come si modificano le prior dei parametri di interesse. Ad ogni iterazione dell'algoritmo (in questo caso 100000), oltre alle stime delle medie e deviazioni standard delle popolazioni, consideriamo anche altre due quantità di interesse per la ricerca che stiamo conducendo, ossia la differenza tra le medie $\mu_y - \mu_x$ e la differenza tra le deviazioni standard $\sigma_y - \sigma_x$ dei nostri gruppi.

Le distribuzioni a posteriori delle sei quantità che ne derivano (le due medie, le due deviazioni standard, la differenza tra le medie e la differenza tra le due deviazioni standard) sono presentate nella figura 1.3. Oltre alla distribuzioni di probabilità, nel grafico sono presentate le stime dei parametri di interesse (media dei valori per le distribuzioni simmetriche e mediana per quelle asimmetriche) e i rispettivi *intervalli di credibilità al 95%* (IC, rappresentati dai segmenti orizzontali; Kruschke, 2015). Proprio perchè fin dall'inizio i parametri sono stati considerati delle variabili casuali, gli intervalli di credibilità esprimono direttamente l'intervallo di valori in cui, con il 95% di probabilità, ricade il valore del parametro incognito, informazione non ricavabile dai tradizionali *intervalli di confidenza* (Hoekstra et al., 2014).

Osservando la figura 1.3 nel suo insieme si può subito intuire la grande mole di informazioni a disposizione. In particolare, osservando il panel f), emerge che le deviazioni standard delle due popolazioni sono con elevata probabilità differenti: la distribuzione a posteriori dei valori plausibili della differenza non è centrata sullo 0, ma piuttosto decentrata verso valori positivi (la mediana è 2.78), e l'intervallo di credibilità non contiene lo 0 (IC 95% = $[0.11, 5.81]$). Per quanto riguarda la differenza tra le medie, panel e), appare più credibile che essa sia positiva piuttosto che nulla o negativa (la media della distribuzione è 2.25); tuttavia l'intervallo di credibilità comprende il valore 0 (IC 95% = $[-1.43, 5.9]$) che pertanto non può essere escluso (o rifiutato). Per avere comunque un'idea dell'efficacia media del trattamento possiamo calcolare la probabilità che la media della popolazione da cui è estratto il campione sperimentale sia maggiore di quella della popolazione da cui proviene il gruppo di controllo: $Pr(\mu_y > \mu_x) = Pr(\mu_y - \mu_x > 0) = 89\%$. Questo valore, che può essere considerato una sorta di dimensione dell'effetto (e che nel caso di popolazioni con medie pressochè uguali dovrebbe assumere valori prossimi al 50%), supporta abbastanza nettamente l'ipotesi che il nuovo metodo abbia

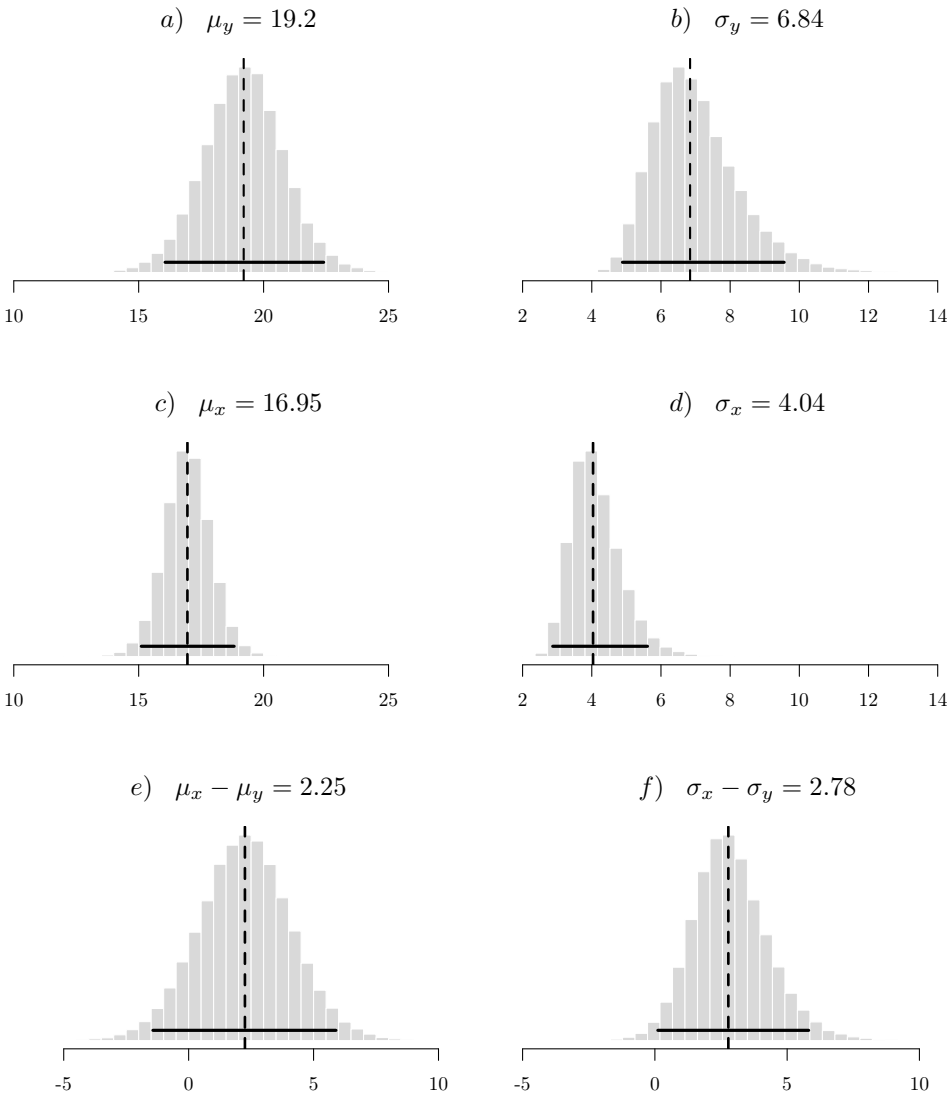


Figura 1.3: Distribuzioni a posteriori dei parametri di interesse. I pedici $y$ e $x$ indicano rispettivamente il gruppo Sperimentale e il gruppo di Controllo. Le linee orizzontali in grassetto indicano gli intervalli di credibilità al 95% per i parametri. Le linee tratteggiate verticali indicano le stime dei parametri (media per le distribuzioni simmetriche, grafici a sinistra, e mediana per le distribuzione asimmetriche, grafici a destra)

complessivamente migliorato l'apprendimento. Dal punto di vista della validità statistica, è importante sottolineare che i risultati sono stati ottenuti senza aver dovuto imporre l'assunzione (dimostratasi poco credibile) di omogeneità delle varianze; assunzione sulla quale si basano sia il t-test che il $d$ di Cohen con cui sono stati inizialmente analizzati questi dati.

In sostanza, dall'analisi bayesiana emerge in primo luogo che il nuovo metodo di insegnamento ha prodotto una maggiore variabilità dei punteggi. Alcuni soggetti hanno molto beneficiato del nuovo metodo, mentre per alcuni il grado di apprendimento si è mantenuto basso. Probabilmente, in linea con recenti approcci che indagano come la variabilità individuale possa interagire con una condizione sperimentale determinando l'esito osservato (si pensi al campo degli interventi), ulteriori variabili di

moderazione andranno indagate (Pluess, 2015) per comprendere cosa funziona, e per chi. Detto ciò si può anche osservare che, in termini di livelli medi di apprendimento, il nuovo metodo sembra essere più efficace di quello tradizionale. Infine, a livello applicativo, i risultati suggeriscono che il nuovo metodo, pur avendo delle buone potenzialità, debba essere rivisto e riprogrammato in modo da estendere i suoi benefici a una maggior proporzione di soggetti.

## Riflessioni conclusive

Negli ultimi anni è cresciuto il dibattito sull'affidabilità delle tecniche di analisi dei dati in psicologia; tecniche di analisi non appropriate (o non aggiornate), tendenza a riportare (e pubblicare) solo dati considerati statisticamente significativi, scarsa attenzione alla rappresentazione grafica delle variabili e dunque ad una analisi dei dati che prenda realmente in considerazione il modo in cui le variabili si distribuiscono, sono tra le critiche mosse con maggior enfasi nel dibattito accademico. A queste critiche ha fatto seguito, costruttivamente, un rinnovato interesse per l'area della buona metodologia e statistica oltre che a livello internazionale (Klugkist et al., 2011; van de Schoot et al., 2013) anche nel nostro Paese, e le criticità sono diventate spunto per il miglioramento ed il dibattito metodologico (Altoè, 2014; Della Sala & Cubelli, 2014; Pastore, 2014; Perugini, 2014).

Come abbiamo discusso nella parte introduttiva di rassegna al dibattito, ogni processo di cambiamento, per passare dalla teoria alla pratica e buona prassi diffusa, richiede tempo, cautela, ed attenzione al dato, per non incorrere in nuovi paradossi. Riprendendo la metafora in apertura, certamente un'automobile adeguata ci è richiesta per muoverci nel traffico attuale, così come adeguate capacità di gestione della stessa perché l'aumento della potenza e delle possibili prestazioni non porti a difficoltà di gestione del mezzo, con incaute o non corrette generalizzazioni che diventano nuovi miti o esiti paradossali. Ci auguriamo con questo lavoro di aver fornito un ulteriore contributo in questa direzione tramite una revisione del dibattito recente e due proposte applicative agevolmente replicabili. A conclusione vorremmo sottolineare come nessuna delle tecniche proposte rappresenti la soluzione tout-court per condurre una buona ricerca. Se i dati sono mal raccolti, mal codificati, non controllati e non esplorati, se non c'è una teoria di riferimento che ispira il lavoro, allora nessuna tecnica da sola conduce a una ricerca affidabile e di buona qualità. Una indispensabile competenza nei modelli teorici indagati, nell'accurata revisione della letteratura, nella progettazione metodologica è fondamentale; con buone analisi possiamo rendere tutti questi sforzi e competenze più evidenti e trovare giovamento in questo modo dal confronto tra discipline in un reciproco arricchirsi di saperi, perché le criticità non siano da blocco ma rappresentino spinte per individiduare ed implementare prassi sempre più rigorose per la ricerca presente e futura.

## Riferimenti bibliografici

Altoè, G. (2014). Approccio bayesiano e replica dei risultati: la dignità dell'ipotesi nulla. *Giornale Italiano di Psicologia*, *41*(1), 55–60.

Altoé, G., & Pastore, M. (2013). L'effetto della numerosità sul significato di un risultato statisticamente significativo. *Giornale Italiano di Psicologia*, *40*(2), 367–376.

American Psychological Association. (2001). *Publication Manual of the American Psychological Association* (5th ed.). Washington, DC: Author.

Anscombe, F. J. (1973). Graphs in statistical analysis. *The American Statistician*, *27*, 17–21.

Bakermans-Kranenburg, M. J., & van IJzendoorn, M. H. (2015). The hidden efficacy of interventions: Gene× environment experiments from a differential susceptibility perspective. *Annual review of psychology*, *66*, 381–409.

Bakker, M., & Wicherts, J. M. (2011). The (mis)reporting of statistical results in psychology journals. *Behavior Research Methods*, *43*(3), 666–678.

Barone, L., Bramante, A., Lionetti, F., & Pastore, M. (2014). Mothers who murdered their child: An attachment-based study on filicide. *Child abuse & neglect*, *38*(9), 1468–1477.

Burnham, K. P., & Anderson, D. R. (2004). Multimodel Inference: Understanding AIC and BIC in Model Selection. *Sociological Methods & Research*, *33*(2), 261–304.

Burnham, K. P., Anderson, D. R., & Huyvaert, K. P. (2011). AIC model selection and multimodel inference in behavioral ecology: some background, observations, and comparisons. *Behavioral Ecology and Sociobiology*, *65*(1), 23–35.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Psychology Press.

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist*, *49*, 997–1003.

Cumming, G. (2012). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. Routledge.

Cumming, G., Fidler, F., Leonard, M., Kalinowski, P., Christiansen, A., Kleinig, A., ... Wilson, S. (2007). Statistical reform in psychology is anything changing? *Psychological Science*, *18*(3), 230–232.

Della Sala, S., & Cubelli, R. (2014). Etica della ricerca e moralità nelle politiche accademiche sono valori da promuovere. *Giornale Italiano di Psicologia*, *41*(1), 81–86.

Fanelli, D. (2012). Negative results are disappearing from most disciplines and countries. *Scientometrics*, *90*(3), 891–904.

Fearon, R., Bakermans-Kranenburg, M. J., Van IJzendoorn, M. H., Lapsley, A.-M., & Roisman, G. I. (2010). The significance of insecure attachment and disorganization in the development of children's externalizing behavior: a meta-analytic study. *Child development*, *81*(2), 435–456.

Fidler, F., Cumming, G., Thomason, N., Pannuzzo, D., Smith, J., Fyffe, P., ... Schmitt, R. (2005). Toward improved statistical reporting in the journal of consulting and clinical psychology. *Journal of consulting and clinical psychology*, *73*(1), 136.

Finch, S., Cumming, G., Williams, J., Palmer, L., Griffith, E., Alders, C., ... Goodman, O. (2004). Reform of statistical inference in psychology: The case of Memory & Cognition. *Behavior Research Methods, Instruments, & Computers*, *36*(2), 312–324.

Fisher, R. A. (1956). *Statistical methods and scientific inference*. Edinburgh, UK: Oliver & Boyd.

Flanagan, O. (2015). *Journal's ban on null hypothesis significance testing: reactions from the statistical arena*. Retrieved from `http://www.statslife.org.uk/opinion/`

Flora, D. B., & Curran, P. J. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological methods*, *9*(4), 466.

Fox, J. (2008). *Applied Regression Analysis and Generalized Linear Models*. SAGE Publications.

Francis, G. (2013). Replication, statistical consistency, and publication bias. *Journal Of Mathematical Psychology*, *57*(5), 153–169.

Gadermann, A. M., Guhn, M., & Zumbo, B. D. (2012). Estimating ordinal reliability for likert-type and ordinal item response data: A conceptual, empirical, and practical guide. *Practical Assessment, Research & Evaluation*, *17*(3), 1–13.

Gelfand, A. E., & Smith, A. F. M. (1990). Sampling-Based Approaches to Calculating Marginal Densities. *Journal of the American Statistical Association*, *85*(410), 398–409.

Gelman, A. (2015). The Connection Between Varying Treatment Effects and the Crisis of Unreplicable Research. A Bayesian Perspective. *Journal of Management*, *41*(2), 632-643.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2014). *Bayesian data analysis* (Vol. 2). Taylor & Francis.

Gelman, A., & Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press.

Geman, S., & Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *6*, 721–741.

Gigerenzer, G., & Marewski, J. N. (2015). Surrogate Science The Idol of a Universal Method for Scientific Inference. *Journal of Management*, *41*(2), 421–440.

Hoekstra, R., Morey, R. D., Rouder, J. N., & Wagenmakers, E.-J. (2014). Robust misinterpretation of confidence intervals. *Psychonomic bulletin & review*, *21*(5), 1157–1164.

Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS medicine*, *2*(8), 696–701.

Jamieson, S. (2004). Likert scales: how to (ab) use them. *Medical education*, *38*(12), 1217–1218.

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological science*, *23*(5), 524-532.

Klugkist, I., van Wesel, F., & Bullens, J. (2011). Do we know what we test and do we test what we want to know? *International Journal of Behavioral Development*, *35*(6), 550–560.

Kruschke, J. K. (2011). Bayesian assessment of null values via parameter estimation and model comparison. *Perspectives on Psychological Science*, *6*(3), 299–312.

Kruschke, J. K. (2015). *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan.* Academic Press.

Kruschke, J. K., & Liddell, T. M. (2017). The bayesian new statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a bayesian perspective. *Psychonomic Bulletin & Review*, 1–29.

Lee, M. D., & Wagenmakers, E.-J. (2014). *Bayesian cognitive modeling: A practical course.* Cambridge University Press.

Lindley, D. V. (2000). The philosophy of statistics. *Journal of the Royal Statistical Society: Series D*(3), 293–337.

Lionetti, F. (2014). What promotes secure attachment in early adoption? The protective roles of infants' temperament and adoptive parents' attachment. *Attachment & human development*, *16*(6), 573–589.

Liu, I., & Agresti, A. (2005). The analysis of ordered categorical data: An overview and a survey of recent developments. *Test*, *14*(1), 1–73.

Lyons-Ruth, K., & Jacobvitz, D. (2008). Attachment disorganization: Genetic factors, parenting contexts, and developmental transformation from infancy to adulthood. In J. Cassidy & P. Shaver (Eds.), *Handbook of attachment.* Guilford Press: Guilford Press.

Masicampo, E. J., & Lalande, D. R. (2012). A peculiar prevalence of p values just below .05. *Quarterly Journal Of Experimental Psychology*, *65*(11), 2271–2279.

Maxwell, S. (2004). The persistence of underpowered studies in psychological research: Causes, consequences, and remedies. *Psychological Methods*, *9*(2), 147–163.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716.

Pace, C. S. (2014). Assessing attachment representations among adoptees during middle childhood and adolescence with the Friend and Family Interview (FFI): clinical and research perspectives. *Frontiers in psychology*, *5*.

Pastore, M. (2014). La significatività della significatività. *Giornale Italiano di Psicologia*, *41*(1), 99–104.

Pastore, M., Nucci, M., & Bobbio, A. (2015). Vita di P: 16 anni di statistche sul GIP. *Giornale Italiano di Psicologia*, *42*, 303–325.

Peeters, M., Zondervan-Zwijnenburg, M., Vink, G., & van de Schoot, R. (2015). How to handle missing data: A comparison of different approaches. *European Journal of Developmental Psychology*, *12*(4), 377–394.

Perugini, M. (2014). La crisi internazionale di credibilità della psicologia come un'opportunità di crescita: Problemi e possibili soluzioni. *Giornale Italiano di Psicologia*, *41*(1), 23–46.

Pluess, M. (2015). Individual Differences in Environmental Sensitivity. *Child Development Perspectives*(doi: 10.1111/cdep.12120).

R Core Team. (2015). R: A Language and Environment for Statistical Computing [Computer software manual]. Vienna, Austria. Retrieved from `http://www.R-project.org/`

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic bulletin & review*, *16*(2), 225–237.

Schimmack, U. (2012). The Ironic Effect of Significant Results on the Credibility of Multiple-Study Articles. *Psychological Methods*, *17*(4), 551–566.

Schwarz, G. E. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6*, 461–464.

Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's Alpha. *Psychometrika*, *74*, 107–120.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant. *Psychological Science*, *22*(11), 1359–1366.

Trafimow, D., & Marks, M. (2015). Editorial. *Basic and Applied Social Psychology*, *37*(1), 1–2.

Tukey, J. W. (1977). *Exploratory data analysis.* Addison-Wesley. Massachusetts, USA: Reading.

van de Schoot, R., Kaplan, D., Denissen, J., Asendorpf, J. B., Neyer, F. J., & van Aken, M. A. (2013). A gentle introduction to bayesian analysis: applications to developmental research. *Child Development*, *85*(3), 842–860.

Yee, T. W. (2010). The VGAM package for categorical data analysis. *Journal of Statistical Software*, *32*(10), 1–34.

Yee, T. W., & Wild, C. J. (1996). Vector Generalized Additive Models. *Journal of Royal Statistical Society, Series B*, *58*(3), 481–493.

Young, N. S., Ioannidis, J. P. A., & Al-Ubaydi, O. (2008). Why Current Publication Practices May Distort Science. *PLOS Medicine*, *5*(10), 1418–1422.

Ziliak, S. T., & McCloskey, D. N. (2008). *The cult of statistical significance.* Ann Arbor, MI: University of Michigan Press.