UNIVERSITÀ DEGLI STUDI DI PADOVA

Dipartimento di Psicologia dello Sviluppo e della Socializzazione

Corso di Laurea Magistrale in
Psicologia Clinica dello Sviluppo

Tesi di Laurea Magistrale

Testare la differenza tra due coefficienti di correlazione: una prospettiva basata sulla Design Analysis

Testing the Difference Between Two Correlation Coefficients: A Design Analysis Approach

Relatore
Prof. Gianmarco Altoè

Laureanda:Allegra Benemerito
Matricola: 2050674

Anno Accademico 2022/2023

A Papà,
a tutto ciò che mi hai lasciato
e a quello che ti sei portato via.
E a Mamma,
per tutto quello che hai sempre fatto per me
e per l'amore che mi dai ogni giorno.
Siete Indispensabili.

INDICE:

Riassunto **1**

Capitolo 1 – La crisi di credibilità in psicologia **3**

1.1 Introduzione: la crisi di credibilità e la sua evoluzione 3
1.2 Riproducibilità e Replicabilità 6
1.3 Perché gli studi non replicano? 8
1.3.1 Questionable Reseach Practices (QRPs) 8
1.3.2 Questionable Measurement Practices e validità 10
1.3.3 Null Hypothesis Significance Testing (NHST) 12
1.4 Come aumentare la replicabilità e superare la crisi? 15
1.5 Scopi della tesi 18

Capitolo 2 – La Design Analysis **19**

2.1 La Design Analysis 19
2.1.1 Il coefficiente di correlazione r di Pearson 21
2.1.2 Errore di Tipo *M*, Errore di Tipo *S* e Potenza 22
2.1.3 Effect size plausibile 24
2.2 Design Analysis Prospettiva e Retrospettiva 25
2.2.1 Design Analysis Prospettiva 26
2.2.2 Design Analysis Retrospettiva 28

Capitolo 3 - Valutare la differenza tra due coefficienti di correlazione via Design Analysis **31**

3.1 Valutare la differenza tra due coefficienti di correlazione 31
3.2 Funzioni in R per la design analysis 32
3.2.1 Introduzione all'utilizzo delle funzioni 33
3.2.2 *retro_rho2* 34
3.2.3 *prosp_rho2* 36
3.3 Approfondimento: relazione tra potenza e differenza di correlazioni 37

Capitolo 4 – Case study: Applicazione della Design Analysis ad un caso reale **40**

4.1 Lo studio originale 40
4.2 Applicazione della Design Analysis al caso 41

4.2.1 Prospective Design Analysis 42
4.2.2 Retrospective Design Analysis 44
Capitolo 5 – Conclusioni 47
BIBLIOGRAFIA 50
Appendice 55
Appendice A: Codici per la Retrospective e Prospective Design Analysis – *retro_rho2* e *prosp_rho2* 55

Riassunto

Nell'arco dell'ultimo decennio sono numerosi gli studi che hanno cercato di replicare procedure di ricerca già esistenti, con l'obiettivo di verificarne la replicabilità. È il caso, ad esempio, dei progetti "Many Labs" e "Reproducibility Project: Psychology" (Open Science Collaboration), i cui risultati sono stati in gran parte deludenti, poiché molti studi si sono rivelati sottopotenziati e gli effetti replicati, quando esistenti, meno robusti degli originali. Ciò ha portato la comunità scientifica a riconoscere di star vivendo una "Crisi di Replicabilità", le cui cause possono essere molteplici.

Un ruolo di rilievo in questa crisi è rivestito, sicuramente, dall'utilizzo di pratiche discutibili messe in atto durante il processo di ricerca e la conseguente analisi statistica: sono numerosi gli studi che, seppur ottenessero risultati statisticamente significativi, erano caratterizzati da una bassa potenza, che rende improbabile ottenere simili effetti.

In questo contesto, la Design Analysis può essere considerata un valido aiuto per quanto riguarda la progettazione e la valutazione di uno studio. Essa consente di ottenere una stima della potenza e dell'Errore di Tipo *S* (segno) e di Tipo *M* (magnitudo), ulteriori rischi inferenziali oltre all'Errore di Tipo I e di Tipo II.

Il presente elaborato di Tesi ha l'obiettivo di analizzare la Crisi di Replicabilità che la psicologia sta attraversando, ponendo enfasi soprattutto sull'analisi statistica, e di proporre come possibile strategia di risoluzione del problema della replicabilità l'impiego della Design Analysis.

Nel primo capitolo verrà, dunque, introdotta la Crisi di Replicabilità, esaminandone i momenti salienti e analizzandone le possibili cause e i possibili rimedi, come l'utilizzo della preregistrazione e di piattaforme open souce.

Il secondo capitolo descriverà la Design Analysis, che permette un'analisi più ampia della potenza, prevedendo l'utilizzo di un effect-size plausibile e la conseguente determinazione di un campione appropriato. Saranno, inoltre, introdotti l'Errore di Tipo *S* e l'Errore di Tipo *M.* Infine, verrà descritta la sua applicazione in maniera prospettiva,

e quindi in fase di pianificazione dello studio, e retrospettiva, da effettuare una volta terminato lo studio.

Il terzo capitolo si soffermerà su come valutare la differenza tra due coefficienti di correlazione, secondo la prospettiva della Design Analysis, con particolare attenzione alle differenze tra coefficienti di correlazione provenienti da campioni indipendenti. Verranno, poi, presentate le due funzioni di R (*retro_rho2* e *prosp_rho2*) implementate per svolgere l'analisi della differenza. Infine, si approfondirà la relazione tra potenza statistica e differenza tra coefficienti di correlazione.

Nel quarto capitolo verrà analizzato l'articolo "Gender differences in reading ability and attitudes: examining where these differences lie" (Logan et al, 2009), prendendo in considerazione una differenza di correlazioni ottenuta da campioni indipendenti e valutata dallo studio come statisticamente significativa.

Il quinto capitolo rappresenta una discussione complessiva di questo elaborato, riassumendo le conclusioni tratte, i limiti delle procedure utilizzate e le possibili prospettive future.

Capitolo 1 – La crisi di credibilità in psicologia

In questo primo capitolo verrà presentata l'evoluzione della crisi di replicabilità che sta attraversando la ricerca in psicologia negli ultimi decenni, ponendo l'accento su riproducibilità e replicabilità, concetti cardine di questa rivoluzione.
Successivamente, verranno analizzate le possibili cause alla base, identificabili nell'utilizzo di Pratiche di Ricerca Discutibili (QRPs), Pratiche di Misura Discutibili (QMPs) e in un utilizzo scorretto dell'analisi statistica, nello specifico del Null Hypothesis Significance Testing (NHST).

Verrà, inoltre, messo un accento sui diversi tipi di validità che contraddistinguono una ricerca, nella fattispecie in psicologia, quali la Validità di costrutto, la Validità interna, la Validità esterna e la Validità statistica delle conclusioni.
Infine, verranno proposte delle possibili risoluzioni per un cambio di rotta, come, ad esempio, l'utilizzo della preregistrazione e di piattaforme open per condividere dati e materiali relativi ad un articolo scientifico

1.1 Introduzione: la crisi di credibilità e la sua evoluzione

La ricerca in ambito psicologico attraversa, da due decenni a questa parte, una crisi causata dalla non replicabilità e riproducibilità di articoli pubblicati (Valentine et al, 2021). Ciò può essere dovuto a bias di pubblicazione (Ferguson et al, 2012), ad un uso e interpretazione scorretto del *p-value* (Ioannidis, 2005), alla messa in atto di quelle che vengono definite Questionable Research Practices (QRPs) (John et al, 2012) e, più in generale, ad un utilizzo scorretto dell'analisi statistica (Pastore et al, 2019).

L'inizio della crisi avviene nel 2005, con la pubblicazione dell'articolo "Why most Published Research Finding Are False" (Ioannidis, 2005), in cui l'autore sostiene che nella ricerca moderna la maggior parte dei risultati ottenuti possa essere falsa. Ioannidis sostiene che una tra le principali motivazioni sia la tendenza a pubblicare risultati basati su conclusioni tratte da studi singoli, sebbene supportate da valori di *p-value* statisticamente significativi ($p\text{-value} < 0.05$). Il *p-value*, però, non può essere l'unico fattore su cui basarsi per stabilire la plausibilità di un risultato (Ioannidis, 2005).

Secondo l'autore, la probabilità che un esito di ricerca sia valido dipende dalla probabilità ante-studio che esso sia valido, dalla potenza (vedere paragrafo 1.3.3 per definizione) dello studio e dal suo livello di significatività statistica.

Egli identifica, inoltre, alcuni fattori che possono influenzare la probabilità che un risultato di ricerca sia plausibile, di seguito riportati in tabella (Tabella 1.1)

▪ Condurre studi di dimensioni ridotte in un campo scientifico (ad esempio, utilizzando un campione piccolo).
▪ Trovare effetti di dimensioni ridotte.
▪ Testare un maggior numero di relazioni e avere una minore selezione delle relazioni testate in un campo scientifico.
▪ Consentire maggiore flessibilità nei disegni, nelle definizioni, nei risultati e nelle modalità analitiche.
▪ Avere maggiori interessi finanziari e altri interessi e pregiudizi all'interno del settore
▪ Essere in un campo scientifico "caldo" (ovvero un ambiente di ricerca attualmente oggetto di grande interesse), con un maggior coinvolgimento di team scientifici

Tabella1.1 Fattori che influenzano la probabilità che un risultato sia valido. Riadattato da Ioannidis (2005)

Altro evento cardine all'interno della storia della crisi è la pubblicazione dell'articolo "Feeling the future: experimental evidence for anomalous retroactive influence on cognition and affect" (Bem, 2011), in cui l'autore sostiene l'esistenza della "pre-cognizione", ovvero la conoscenza o la sensazione di un evento futuro, senza che questo possa essere anticipato attraverso processi noti. Nel 2012 Galak e collaboratori portarono a termine 7 esperimenti che dimostrarono l'infondatezza alla base di questa teoria, confermata poi da altre replicazioni avvenute nel 2023.

All'interno del quadro della crisi di credibilità è impossibile, inoltre, non citare la frode commessa da Diedrich Stapel, nel 2011. In uno dei suoi studi, egli dimostrò come ambienti disordinati potessero essere facilitatori della discriminazione razziale (Stapel, 2011). Alcuni anni dopo, egli stesso ammise di aver falsificato dati di dozzine di articoli e di almeno 14 su 21 tesi di dottorato (Stapel, 2014).

Nel 2014 venne pubblicato il primo tentativo di replicazione in larga scala, progetto che prende il nome di "Many Labs" e che testò la replicabilità di 13 risultati di

esperimenti psicologici in 12 paesi, coinvolgendo 6344 partecipanti. I risultati sembravano sorprendenti, dal momento che il 77% degli studi (10/13) replicò con successo. Alcuni ricercatori contestarono, però, che molti degli esperimenti presi in considerazione fossero già pensati per essere replicabili (Pennington, 2023).

Sulla scia del Many Labs Replication Project, nel 2015 l'Open Science Collaboration (OSC) attuò il "Reproducibility Project: Psychology" (Open Science Collaboration, 2015), in cui tentò di replicare 100 risultati scelti casualmente tra le riviste di psicologia più influenti. Solo il 36% degli studi presi in considerazione replicò con successo (p-value < 0.05) e, anche tra questi, l'effect size diminuì di circa la metà rispetto agli originali.

Con i risultati dell'OSC, la discussione sulla replicabilità iniziò ad espandersi anche ad altri campi della scienza. Nel 2016, Baker coinvolse 1500 scienziati per investigare questo tema anche in altri ambiti scientifici: approssimativamente il 90% di loro convennero sull'esistenza di una crisi. Il 40% dichiarò di non essere in grado di replicare propri esperimenti e più del 60% anche esperimenti di altri scienziati (Baker, 2016).

Secondo alcuni autori, comunque, questo momento di "crisi" che stanno attraversando la psicologia e le altre scienze potrebbe essere ottimisticamente visto come una rivoluzione (Munafo'et al, 2017; Fanelli 2018; Vazire, 2018; Munafo' et al, 2022), considerando i progressi che si stanno facendo e che si possono fare per invertire la tendenza.

In figura 1.1 un breve riassunto dei momenti salienti che hanno caratterizzato la crisi fino ad oggi.

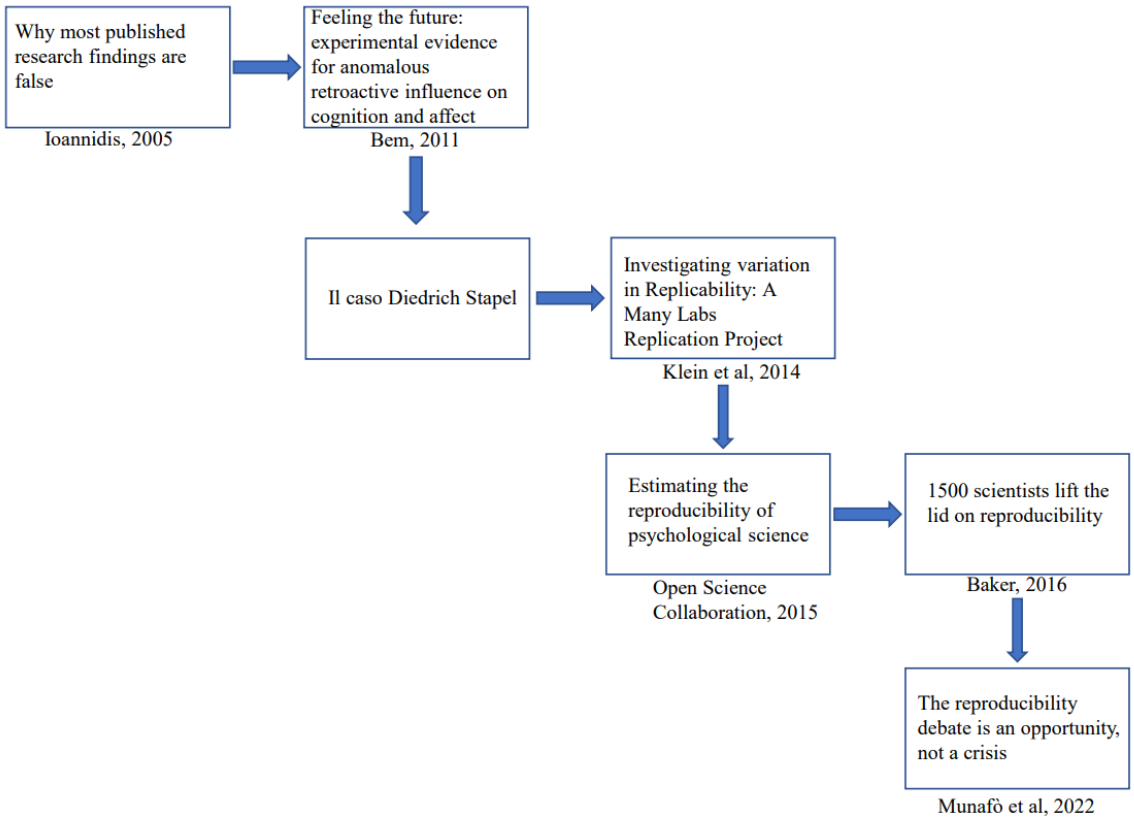

Figura1.1, Schema dei punti salienti della Crisi di Replicabilità. Riadattato da Pennington (2023)

1.2 Riproducibilità e Replicabilità

Per comprendere a pieno la portata della Crisi di credibilità bisogna comprendere i due concetti che ne sono alla base, la riproducibilità e la replicabilità.

La riproducibilità si riferisce alla capacità di ottenere risultati coerenti e simili quando un esperimento o uno studio viene ripetuto utilizzando gli stessi metodi, protocolli e condizioni.

Tutti i risultati riportati da uno studio devono essere, innanzitutto, riproducibili. Ciò implica che ricerche successive devono ottenere risultati significativi e coerenti con l'originale, utilizzando gli stessi dati e procedure (Nosek et al, 2022).

I test di riproducibilità possono fallire, sostanzialmente, per due motivazioni principali (Nosek et al, 2022):

1. L’analisi originale non può essere ripetuta per indisponibilità di dati o di informazioni indispensabili allo studio.
2. La rianalisi ottiene risultati diversi dallo studio pilota. Ciò può avvenire a causa di errori sia nello studio originale sia in quello di riproducibilità.

La riproducibilità è alla base della credibilità di un’evidenza scientifica ma non è sufficiente, ed è qui che entra in gioco la replicabilità, considerata da alcuni *conditio sine qua non* per rendere scientifico un risultato empirico (Schmidt, 2009).

Per replicabilità si intende ripetere la procedura di uno studio utilizzando nuovi dati per provare se i nuovi risultati ottenuti sono coerenti con quelli dello studio iniziale (Nosek et al, 2022). Essa è alla base della scienza: un effetto osservato può essere effettivamente considerato una scoperta solo nel momento in cui sia stato replicato molteplici volte da gruppi di ricerca differenti (Allen et al, 2023).

Bisogna distinguere tra due tipi di replicazione, quella diretta (o esatta) e quella concettuale. Quando uno studio ripete lo stesso disegno sperimentale e i metodi di una ricerca il più fedelmente possibile, con il fine di ottenere gli stessi risultati, si è di fronte ad una replicazione diretta. Essa non va confusa con quella concettuale, che si pone come scopo quello di cercare lo stesso effetto dello studio originale, usando metodi differenti e in condizioni diverse (Derksen & Morawski, 2022).

Le repliche sono importanti poiché consentono di stabilire la generalizzabilità di un effetto e aiutano ad identificare falsi positivi (Nosek, 2022).

Nelle repliche dirette, la non replicabilità dei dati originali potrebbe stare a significare che qualcosa nell’impostazione originale potrebbe aver influenzato i risultati che non possono, dunque, essere generalizzabili oltre alle condizioni specifiche dello studio iniziale. Allo stesso modo, se non emergono risultati statisticamente significativi nelle nuove condizioni, ciò potrebbe significare che i risultati trovati nel setting originale non possono essere generalizzati a quello nuovo (o che l’effetto originale non sia affidabile) (Allen et al, 2023)

1.3 Perché gli studi non replicano?

Consultando la letteratura in merito risulta evidente che le cause alla base della crisi di credibilità siano molteplici (Pennington, 2023).

Diversi autori si sono concentrati su come i ricercatori aumentino sistematicamente i tassi di Errore di Tipo I (che consiste nel rifiutare l'ipotesi nulla quando essa è vera) e esagerino la dimensione dell'effetto ottenuto dai propri studi (Shrout & Rodgers, 2018). Ciò è spesso incoraggiato dal contesto in cui il ricercatore lavora, all'interno del quale assunzioni, incarichi, promozioni e sovvenzioni sono strettamente legate al numero di pubblicazioni in riviste rilevanti (Nosek et al, 2012)

A tal riguardo non si possono non citare i bias di pubblicazione, che consistono nel pubblicare esclusivamente studi che ottengano esiti statisticamente significativi (Ferguson & Brannick, 2011). Studi pubblicati in un campo di ricerca non rappresentativo dell'effettiva popolazione di dati possono tradursi in interpretazioni distorte delle ricerche sia per quanto riguarda le revisioni narrative che le meta-analitiche (Ferguson & Brannick, 2011).

La pressione a pubblicare può portare i ricercatori a trascurare il metodo scientifico in favore della messa in atto di strategie che si rivelano controproducenti per la ricerca, quali le *Questionable Research Practices* (QRPs)

1.3.1 Questionable Reseach Practices (QRPs)

Con Questionable Research Practices ci si riferisce ad un termine ombrello che racchiude i comportamenti che i ricercatori operano – volontariamente o meno – per modificare i propri risultati (Parsons et al, 2022). Esse possono essere considerate come la via di mezzo tra le condotte responsabili di ricerca (Responsible conduct of research, RCR) e quelle di fabbricazione, falsificazione e plagio (FFP) (Steneck, 2006). Banks et al (2016) le definiscono, esattamente, come "pratiche di progettazione, analisi o reporting discutibili per il loro potenziale impiego allo scopo di presentare prove distorte a favore di un'asserzione" e secondo Bouter et al (2016) esse coinvolgono diverse azioni relative alla fase di progettazione dello studio, al processo di raccolta e rendicontazione dei dati e alla collaborazione in ricerca.

Si possono distinguere diversi tipi di QRPs, molti dei quali accostabili alla pratica del p-hacking (o *data dredging*), che consiste nella manipolazione dei dati al fine di

ottenere risultati statisticamente significativi (Simmons et al, 2011). Ciò può avvenire, per esempio, tramite l'esclusione degli outliers dai dati, l'approssimazione scorretta del *p-value*, per fare in modo che esso risulti minore di 0.05 (es. $p = 0.058$ approssimato a $p = 0.05$) o la misurazione della stessa variabile dipendente attuata molteplici volte in maniere diverse. (Pennington, 2023). Un altro modo in cui può esprimersi il p-hacking è, ad esempio, tramite la pratica dell'*optional stopping*, che prevede un'analisi ripetuta dei dati durante la fase di raccolta, che procede finché non si raggiunge il valore di p-value desiderato (Simmons et al, 2011).

Un'altra pratica discutibile messa in atto nel campo della ricerca è costituita dall'HARKing (Hypothesizing After Results are Known), che si verifica nel momento in cui il ricercatore formula ipotesi successivamente all'analisi dei dati sviluppando, quindi, una teoria derivante dai risultati ottenuti (Banks et al, 2016). Questa pratica risulta problematica poiché la ricerca esplorativa, inizialmente formulata, viene presentata come confermativa, facendo in modo tale che il ricercatore confermi sempre la propria ipotesi tramite i risultati che ottiene (Rubin, 2017). Tale comportamento va contro il modello ipotetico-deduttivo della scienza (Chambers, 2017), in cui la fase della costruzione delle ipotesi precede quella della raccolta dati, che viene poi utilizzata per confermare la teoria iniziale (Figura 1.2).

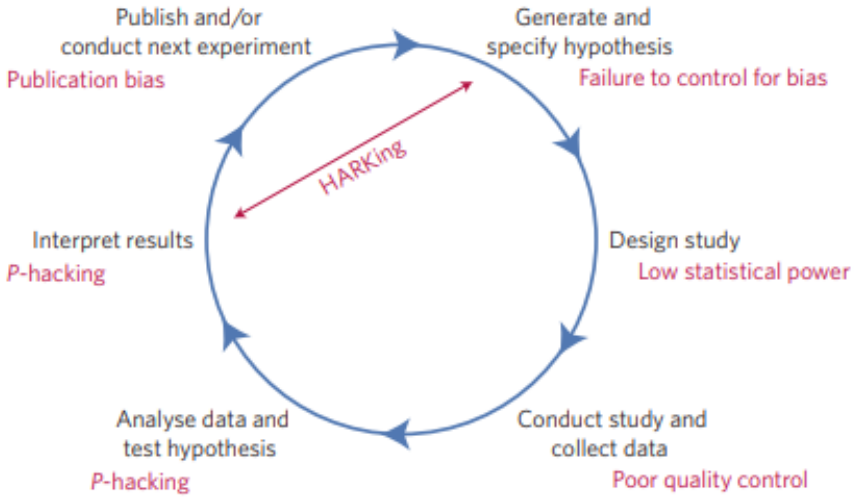

Figura1.2: da "A manifesto for reproducible science" (Munafo' et al, 2017). Rappresentazione del modello ipotetico-deduttivo e delle diverse QRPs riscontrabili in tale processo

Simmons e colleghi (2011) suggeriscono alcune accortezze da mettere in atto per arginare questo problema (Tabella 1.2)

1. Stabilire i criteri per terminare la raccolta dati prima che essa inizi e riportarli all'interno dell'articolo
2. Raccogliere almeno 20 osservazioni o fornire giustificazioni in merito al costo della raccolta dati
3. Elencare tutte le variabili considerate nello studio
4. Riportare tutte le condizioni sperimentali, incluse le manipolazioni fallite
5. Se vengono eliminati dati, riferire anche i risultati dell'analisi statistica che includono tali dati
6. Se un'analisi include una covariata, includere i risultati statistici dell'analisi escludendo tale covariata

Tabella 1.2. Suggerimenti per autori. *Nota.* Adattata da Simmons et al (2011)

1.3.2 Questionable Measurement Practices e validità

Una delle decisioni che il ricercatore deve prendere durante una ricerca riguarda il come misurare il costrutto psicologico interessato. In psicologia, però, i possibili approcci per la valutazione sono numerosi e ciò implica che il ricercatore si trovi davanti svariate alternative tra cui scegliere.

Da una review di Flake e colleghi (2020) emerge che in numerosi studi in psicologia non sia chiaro se le misure utilizzate siano valide per il costrutto in analisi. Spesso, infatti, non viene specificato quali costrutti siano presi in considerazione, come vengano misurati e perché sia stata utilizzata una specifica metodologia piuttosto che un'altra. Questa mancanza di trasparenza all'interno della ricerca può essere accostata a quelle che vengono definite Pratiche di misura discutibili (Questionable Measurement Practices, QMPs), che si riferiscono alle decisioni prese dai ricercatore che installano il dubbio riguardo alla validità delle misure utilizzate nello studio e alla validità delle conclusioni dello stesso. Se le misure non sono valide non possono esserlo neppure i risultati e le conclusioni ottenute dallo studio stesso e, di conseguenza, nel momento in cui esso viene replicato, la replica è destinata a fallire in principio.

Shadis et al (2002) definisce quattro tipi di validità che contribuiscono alla validità generale delle conclusioni:

- Validità interna
- Validità esterna
- Validità di costrutto
- Validità delle conclusioni statistiche

La validità interna si riferisce al modo in cui uno studio stabilisce una relazione di causa-effetto tra le variabili indipendenti e dipendenti (Cook & Campbell, 1979). Se non ci sono le informazioni necessarie per determinare se le proprietà delle misure di uno studio differiscano tra due differenti condizioni di trattamento o nel tempo, la validità interna non può essere completamente valutata.

La validità esterna riguarda la possibilità di generalizzare i risultati ottenuti a popolazioni diverse e in setting diversi (Cook & Campbell, 1979). Ciò può influenzare negativamente i tentativi di replica, nel momento in cui essi vengano condotti su un campione di cultura differente o se i materiali utilizzati siano stati tradotti in un'altra lingua e risultino poco comprensibili per il nuovo campione (Pennington, 2023)

La validità di costrutto si riferisce al modo in cui i costrutti in uno studio vengono operazionalizzati e, infine, la validità delle conclusioni statistiche riguarda la correttezza delle conclusioni dell'analisi statistica.

Flake & Fried (2020) suggeriscono, nel loro articolo "Measurement Schmeasurement", alcune domande a cui il ricercatore dovrebbe rispondere, durante il proprio studio, con il fine di identificare i possibili QMP in cui potrebbero incorrere ed evitarli (Tabella 1.3). Rispondere a queste domande può essere utile per la valutazione della validità della ricerca e consente degli studi di replicabilità significativi.

Domanda	Obiettivo
1. Qual è il costrutto? 2. Come e perché sono state selezionate le misure? 3. Quali misure sono state utilizzate per operazionalizzare il costrutto? 4. Come si quantificano le misure? 5. Le scale sono state modificate? Se si, perché? 6. Sono state create nuove misure?	1. Definire il costrutto e le teorie a sostegno di esso 2. Giustificare la scelta delle misure considerate, riportando prove sulla validità 3. Descrivere le misure e confrontarle con il costrutto 4. Descrivere la codifica delle risposte; riportare gli stimoli inclusi in ciascun punteggio; descrivere i calcoli e le analisi condotte 5. Descrivere eventuali modifiche apportate, giustificandole 6. Motivare l'uso di nuove misure, riportando tutte le prove di validità disponibili

Tabella 1.3. Domande utili da porsi per evitare di incorrere in QRPs. Riadattato da Flake & Fried (2020)

1.3.3 Null Hypothesis Significance Testing (NHST)

Diversi articoli in letteratura evidenziano come altra criticità nella ricerca in psicologia anche l'utilizzo e l'eccessiva dipendenza dal Null Hypothesis Significance Testing (NHST) (Gigerenzer et al., 2004; Gelman, 2018; McShane et al., 2019)

Gigerenzer (2004), descrive in questo modo i 3 passaggi principali:

1. Impostare un'ipotesi nulla (= nessuna differenza media/nessuna correlazione) senza specificare le ipotesi di ricerca
2. Usare un livello convenzionale ($\alpha_{critico}$ o livello di significatività critico) pari a .05 (5%). Se il p-value osservato è inferiore all'$\alpha_{critico}$ è possibile accettare l'ipotesi di ricerca)
3. Eseguire sempre la procedura in questo modo

Questo approccio risulta una combinazione delle procedure proposte originariamente da Fisher e Neyman & Pearson senza, però, che nessuna delle due venga coerentemente applicata: viene, infatti, utilizzato il livello di significatività appartenente all'approccio di Fisher con una strategia simile a quello di Neyman & Pearson, che prevede di scegliere tra l'ipotesi nulla H_0 e l'ipotesi alternativa H_1, senza che quest'ultima venga formalizzata o presa in considerazione; come conseguenza di questa mancanza, l'Errore di secondo tipo (β) e la potenza (1-β) possono essere ignorati.

In questo approccio, inoltre, non essendo stata formulata alcuna ipotesi alternativa, non si può "accettare l'ipotesi nulla" ma "rifiutare o non poter rifiutare" H_0, decisione che viene presa tramite l'analisi del *p-value*, valutato rispetto ad un $\alpha_{critico}$ fissato per convenzione a 0.05. Se il p-value è minore di 0.05 si può rifiutare l'ipotesi nulla; se, al contrario, il *p-value* è maggiore di α, H_0 non potrà essere rifiutata (Figura 1.3)

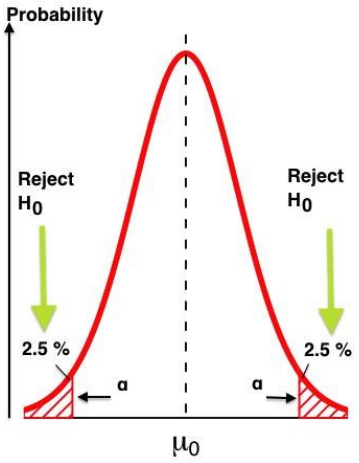

Figura 1.3: Null Hypothesis Significance Testing (NHST). Distribuzione campionaria della statistica del test sotto l'ipotesi nulla H_0. α è fissato a 0.05, se p-value < 0.05 □ è possibile rifiutare H_0. Figura tratta da "Are Your P-values Killing your AB Testing Efforts?" (https://www.johnquarto.com/2014/09/are-your-p-values-killing-your-ab-testing-efforts/)

L'NHST si concentra, dunque, sull'uso e sulla valutazione del p-value per stabilire se i risultati di uno studio siano statisticamente significativi o meno. Questa procedura è stata fortemente criticata, soprattutto perché non prende in considerazione aspetti molto rilevanti per la ricerca, quali la potenza e l'effect size (Gigerenzer, 2018).

La potenza (1-β) di un test statistico rappresenta la probabilità di rifiutare correttamente l'ipotesi nulla (H0) quando questa è effettivamente falsa, oppure di

accettare correttamente l'ipotesi alternativa quando questa è vera. In altre parole, la potenza riflette la capacità del test di individuare un effetto esistente nel campione.

Ignorare la potenza in uno studio può comportare la realizzazione di indagini con una bassa potenza statistica, il che significa che vi è una scarsa probabilità di rilevare un effetto anche se esiste realmente. Se la potenza è bassa, anche se l'effetto è presente, ci sono basse probabilità che esso si riesca a rilevare tramite il test statistico utilizzato.

Inoltre, nel caso in cui un ricercatore dovesse trovare un effetto nonostante una potenza statistica bassa, egli potrebbe attribuirgli ancora più importanza, perché il ritrovamento è considerato "statisticamente significativo". Questo fenomeno è noto come "What does not kill statistical significance makes it stronger" (Loken & Gelman, 2012). Tuttavia, è importante sottolineare che questa interpretazione può essere fuorviante. La significatività statistica da sola non può fornire una valutazione completa dell'importanza pratica o della dimensione dell'effetto.

È fondamentale considerare la potenza sin dall'inizio della pianificazione dello studio, al fine di garantire che la ricerca sia in grado di rilevare effetti di interesse con un'adeguata probabilità. Inoltre, è necessario considerare tutti i fattori che possono influenzare la potenza, come la dimensione del campione, gli errori di misura, i gradi di libertà del ricercatore e altri fattori che possono contribuire al ritrovamento di effetti minimi considerati statisticamente significativi, ma che potrebbero non avere una rilevanza pratica significativa. Quando si parla di replicabilità, prendere in considerazione la potenza statistica è importante per due motivazioni: la prima è che se l'ipotesi alternativa è corretta, ma la potenza è bassa, le possibilità di replicare risultati significativi sono basse; la seconda è che, nel caso in cui la potenza sia bassa, i risultati potrebbero sopravvalutare l'effect size dell'effetto studiato (Button et al, 2013).

Nella ricerca psicologica, l'effect size rappresenta l'entità dell'effetto o della relazione tra le variabili studiate. Costituisce una misura oggettiva dell'importanza o della dimensione pratica degli effetti osservati, indipendentemente dalla dimensione del campione o dalla significatività statistica. Non considerarlo nel contesto di ricerca può avere diverse implicazioni per la replicabilità degli studi, per quanto riguarda l'interpretazione degli effetti riscontrati, l'affidabilità dei risultati, la potenza e le meta analisi dello studio.

1.4 Come aumentare la replicabilità e superare la crisi?

Come risposta alla crisi di replicabilità, molti psicologi si sono prodigati per lo sviluppo e la promozione di tecniche di Open Science, come la preregistrazione, che implica la condivisione pubblica di tutti i dettagli riguardanti lo studio, quali disegno di ricerca, ipotesi e risultati, prima ancora che siano stati raccolti e analizzati i dati (Nosek, 2018), con lo scopo di fornire una presentazione trasparente delle decisioni e delle modifiche apportate fino alla pubblicazione dell'articolo (Spitzer & Mueller, 2023).

L'utilizzo delle preregistrazioni in psicologia è in continuo incremento: ad esempio, il numero di preregistrazioni contenute nell'Open Science Framework (OSC), vede un raddoppiamento annuo dal 2012 al 2017 (Nosek et al, 2018) e, come mostrato da uno studio condotto nel 2018, il 44% dei ricercatori in psicologia intervistati hanno confermato di aver preregistrato ipotesi o analisi fino all'anno prima (Christensen et al, 2020). La stessa crescita esponenziale è evidente anche per quanto riguarda un'altra piattaforma di condivisione, AsPredicted, in cui vengono caricati circa 1200 nuovi studi al mese e, ad oggi, sono presenti più di 80000 preregistrazioni condivise da circa 2100 istituzioni (https://credlab.wharton.upenn.edu/, dati in continuo aggiornamento).

Un'altra forma di preregistrazione è rappresentata dai Registered Reports (RR), che consentono di ottenere peer reviews in uno stadio precedente alla pubblicazione della ricerca. I RR si articolano in due stadi (Figura 1.4):

- Nello Stadio 1, i ricercatori inviano alla rivista scientifica una proposta di articolo che includa introduzione, metodo e progetto d'analisi. Successivamente, viene fornita una revisione con annessi suggerimenti e modifiche da apportare per migliorare lo studio prima che avvengano raccolta e analisi dei dati. Una volta apportate le migliorie del caso, i ricercatori potrebbero ottenere l'IPA (In pinciple acceptance), tramite la quale la rivista accetta il lavoro svolto nel primo stadio e si impegna a pubblicare l'articolo indipendentemente dai risultati da esso ottenuti. In questo stadio avviene automaticamente anche la preregistrazione dell'articolo;
- In seguito, avvengono la raccolta e l'analisi dei dati e l'invio dell'articolo, che unisce alla proposta iniziale anche i risultati e la discussione. A questo punto, avviene una seconda revisione in cui viene appurato che il ricercatore abbia seguito il progetto iniziale (o che abbia appropriatamente indicato qualsiasi modifica) e che i risultati siano stati accuratamente interpretati. Infine, l'articolo

viene accettato. Va sottolineato, inoltre, che arrivati a questo punto la revisione non possa riguardare ciò che è stato accordato nella fase precedente.
(Chambers & Tzavella, 2021)

Figura 1.4, modello di pubblicazione dei Registered Reports, tratto da "Registered Reports: Peer review before results are known to align scientific values and practices" (Center for Open Science, 2019)

Quali sono, dunque, i benefici dei Registered Reports in relazione alla crisi di replicabilità?

Prima di tutto, i criteri di accettazione dei RR sono più severi rispetto a quelli di molte riviste, dato che i ricercatori devono presentare ricerche con adeguati design e analisi statistiche. In questo modo, gli articoli pubblicati sono caratterizzati da una potenza più elevata, che riduce le probabilità di commettere errori di primo tipo (ovvero falsi positivi) e di secondo tipo (falsi negativi).

L'utilizzo dei RR pone l'attenzione alle metodologie utilizzate, più che ai risultati, portando i ricercatori ad aggiungere ulteriori test a sostegno delle loro ipotesi (Lakens et al, 2018). Ciò consente di spostare il focus dal p-value che, come detto in precedenza, non è sufficiente per garantire la significatività dei risultati.

I Registered Reports, inoltre, potrebbero facilitare la replicabilità e la riproducibilità degli articoli anche perché, tramite tutti i controlli ai quali essi sono sottoposti, risulta più facile rilevare le possibili QRPs. Essi, inoltre, potrebbero mitigare anche la messa in atto di bias di pubblicazione poiché l'articolo verrebbe pubblicato indipendentemente dai risultati da esso ottenuti. (Pennington, 2023).

Al fine di migliorare la conduzione della ricerca, inoltre, Cumming (2013) propone, nel suo articolo "The New Statistics: Why and How", 25 linee guida da seguire (Tabella 1.4), soprattutto nel caso in cui venisse utilizzato l'approccio NHST.

1. Promuovere l'integrità della ricerca
2. Collaborare con altri ricercatori
3. Segnalare ogni dettagli dello studio
4. Chiarire la tipologia di ogni risultato – se speculativo o meno
5. Effettuare studi di replicazione
6. Costruire una disciplina quantitativa cumulativa
7. Se possibile, evitare il pensiero dicotomico (tipico dell'NHST)
8. Ricordare che i risultati ottenuti sono una possibilità su una sequenza infinita
9. Non affidarsi ciecamente a nessun p-value
10. Se possibile, evitare di utilizzare l'NHST
11. Superare l'NHST e utilizzare metodi più appropriati
12. Interpretare gli effect size con giudizio
13. Interpretare il proprio intervallo di confidenza (IC), tenendo presente che potrebbe non essere rappresentativo
14. Preferire IC al 95% rispetto (???)
15. Se l'effect size di interesse è una differenza, utilizzare l'intervallo di confidenza su quella differenza per l'interpretazione
16. Considerare l'interpretazione dell'effect size e degli intervalli di confidenza come modo efficace per analizzare i risultati del proprio studio
17. Quando appropriato, utilizzare gli intervalli di confidenza sulle correlazioni
18. Utlizzare meta-analisi
19. Utilizzare un modello a effetti casuali (?) per la meta-analisi
20. Pubblicare i risultati in modo da facilitarne l'inclusione in future meta-analisi
21. Sforzarsi per aumentare l'informatività della ricerca
22. Se si utilizza l'NHST, prendere in considerazione di calcolare la potenza
23. Fare attenzione ad ogni dichiarazione di potenza che non indici un effect-size
24. Utilizzare un'analisi di precisione per la pianificazione, quando utile
25. Adottare una prospettiva di stima quando si considerano questioni di integrità della ricerca

Tabella 1.4. Linee guida di ricerca da seguire. Riadattato da Cumming (2013)

Per quanto riguarda, invece, la risoluzione della crisi da un punto di vista più statistico, un possibile ausilio potrebbe essere identificato nella Design Analysis (DA), che verrà descritta nel dettaglio nel capitolo seguente.

1.5 Scopi della tesi

In questo capitolo è stata descritta la Crisi di credibilità che la psicologia, assieme ad altre scienze, sta attraversando negli ultimi decenni, mettendo a fuoco alcuni dei principali motivi che hanno ad essa contribuito (QRPs, bias, NHST). Dato il basso tasso di replicabilità degli studi, sono stati identificati alcuni rimedi per migliorare vari aspetti del processo di ricerca, come la pre-registrazione e l'utilizzo dell'Open Science Framework.

L'obiettivo di questa tesi è proporre e descrivere la Design Analysis (DA), un nuovo metodo statistico basato sulla definizione di un effect-size plausibile, ricavato a seguito di uno studio approfondito della letteratura. La DA rappresenta un'innovazione nel campo della ricerca e il suo beneficio principale è costituito dal fatto che per la sua applicazione in ricercatore è portato a riflettere minuziosamente sulla ricerca che sta sviluppando. La Design Analysis, inoltre, consente di ottenere informazioni fondamentali per lo studio, sia nel caso di progettazione che in quello di valutazione.

Nei capitoli seguenti verrà, dunque, analizzata la Design Analysis in entrambe le sue forme, prospettiva e retrospettiva, e verranno fornite due funzioni, *prosp_rho2* e *retro_rho2*, formulate per utilizzare la DA nel caso della valutazione di differenze tra coefficienti di correlazione ricavate da campioni indipendenti, che saranno poi utilizzate nel Capitolo 4 per uno studio di un caso reale, "*Gender differences in reading ability and attitudes: examining where these differences lie*"

Capitolo 2 – La Design Analysis

In questo capitolo verrà descritta la Design Analysis (DA), un approccio statistico proposto da Gelman & Carlin (2014) che porta ad un'analisi più appropriata della ricerca. Essa implica la determinazione di una dimensione campionaria appropriata, stimata considerando gli errori di Tipo 1 e Tipo 2 e un "effect size plausibile". Con la DA, inoltre, sono stati introdotti altri due rischi inferenziali, costituiti dall'Errore di tipo *M* (magnitudo) e l'Errore di tipo *S* (segno).

Nel corso del secondo capitolo verranno, inoltre, descritti i due tipi di Design Analysis applicabili: la DA prospettiva e la DA retrospettiva.

2.1 La Design Analysis

Come messo in evidenza nel corso del primo capitolo, una delle cause che ha portato alla Crisi di Replicabilità vissuta dalla psicologia è identificabile nell'uso scorretto e improprio dell'analisi statistica, che porta a conseguenze come studi con scarsa potenza e non replicabili.

È per questo motivo che Gelman e Carlin (2014) hanno introdotto la "*Prospective & Retrospective Design Analysis*", un'analisi più ampia della potenza, che prevede la determinazione di un campione appropriato, con livelli predefiniti di errori di Tipo I e di tipo II e un "effect-size plausibile" (Gigerenzer et al, 2004). In particolare, la Design Analysis considera esplicitamente anche altri due tipi di rischi inferenziali, quali l'Errore di Tipo *M*, che si riferisce alla magnitudo dell'effetto, e l'errore di Tipo *S*, che riguarda il segno dell'effetto riscontrato (Gelman & Carlin, 2014).

Idealmente, la Design Analysis dovrebbe essere messa in atto durante la fase preliminare di uno studio, nel momento in cui esso viene progettato. In questa fase, essa prende il nome di "Prospective Design Analysis" e può essere una risorsa poiché consente di identificare una dimensione campionaria relativa ad una buona potenza, tenendo in considerazione anche i due errori sopra citati. Questo non è, però, l'unico contesto in cui la Design Analysis può essere utile: essa può essere, infatti, sfruttata anche per valutare studi già condotti e in cui il disegno di ricerca è già noto, supportando o meno i risultati ottenuti dalla ricerca tramite la valutazione degli errori di Tipo *M* ed *S*. In questo caso,

ciò che viene effettuato è una Retrospective Design Analysis, da non confondere con l'analisi della potenza post-hoc: nel primo caso, viene definito un effect-size plausibile riferendosi alla letteratura presente o ad altre informazioni esterne allo studio preso in considerazione; nel secondo, invece, l'effect-size plausibile viene identificato basandosi su risultati ottenuti nello studio (Altoè et al, 2020; Bertoldo et al, 2022). Indipendentemente dal tipo di DA utilizzata, la relazione tra Errore di Tipo *M*, Errore di Tipo *S* e potenza rimane invariata (Altoè et al, 2020).

Secondo la prospettiva di Gelman & Carlin, la Design Analysis viene applicata a misure dell'effetto non standardizzate. Tale applicazione risulta poco pratica in ambito psicologico, poiché raramente si utilizzano misure di questo tipo per lo studio dei fenomeni. Altoè et al (2020) e Bertoldo et al (2022) hanno proposto, perciò, due adattamenti della DA utilizzando misure alternative di valutazione dell'effect-size, quali la d di Cohen e il coefficiente di correlazione r di Pearson.

Partendo da queste premesse, Altoè et al (2020) propongono una procedura basata sulle simulazioni di Monte Carlo per realizzare una DA. Tale procedura può essere riassunta nei seguenti tre passi:

1. Identificare un effect-size plausibile per lo studio, basandosi su un'ampia revisione della letteratura e/o su meta-analisi e/o considerazioni teoriche.
2. Sulla base del disegno sperimentale di interesse, eseguire poi un ampio numero di simulazioni (es. 100000) in base all'effect size plausibile identificato. Questa procedura ha lo scopo di fornire informazioni su cosa aspettarsi nel momento in cui lo studio fosse replicato un numero infinito di volte, assumendo che l'effect-size precedentemente identificato sia valido.
3. Dato un livello fisso di errore di Tipo I (es. 0.05), calcolare la potenza, l'errore di Tipo *M* e di Tipo *S*. Nello specifico:
    - Potenza: rapporto tra il numero di risultati significativi ottenuti e il numero di replicazioni
    - Errore di Tipo *M*: rapporto tra la media dei valori assoluti degli effect-size replicati, statisticamente significativi e l'effect-size plausibile
    - Errore di Tipo *S*: rapporto tra il numero di risultati significativi con segno opposto rispetto all'effect-size plausibile e il numero totale di risultati significativi.

Questa procedura può essere messa in atto tramite il pacchetto di R PRDA, per la Prospective e Retrospective Design Analysis, disponibile su CRAN (https://CRAN.R-project.org/package=PRDA). Tale funzione, oltre a lavorare sugli effetti precedentemente citati, presenta ulteriori applicazioni legate alla DA.

2.1.1 Il coefficiente di correlazione r di Pearson

Il coefficiente di correlazione di Pearson è una misura standardizzata dell'effect-size che indica la forza e la direzione di una eventuale relazione di linearità tra due variabili continue (Cohen, 1988; Ellis, 2010).

Date due variabili X e Y, l'indice r di Pearson è definito come il rapporto tra la covarianza[1] tra le due variabili e il prodotto delle loro deviazioni standard:

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

Il coefficiente ρ può assumere valori compresi tra ± 1. Nello specifico:

- $\rho_{XY} > 0$: le variabili considerate sono correlate positivamente
- $\rho_{XY} = 0$: le variabili non sono correlate
- $\rho_{XY} < 0$: le variabili sono correlate negativamente

A seconda del valore assunto, inoltre, può essere stabilita anche l'intensità della correlazione che intercorre tra le due variabili studiate (Cohen,1988):

- Se $|\rho_{XY}| = 0.1$ la correlazione tra le due variabili è debole
- Se $|\rho_{XY}| = 0.3$ la correlazione tra le due variabili è moderata
- Se $|\rho_{XY}| = 0.5$ la correlazione tra le due variabili è forte

I valori sopra riportati sono stati indicati da Cohen (1988) come valori di riferimento da utilizzare nel momento in cui il ricercatore non possiede altre informazioni utili per trarre conclusioni. In caso contrario, l'effect-size riscontrato va sempre contestualizzato al contesto di ricerca in cui lo studioso si pone.

[1] La covarianza consente di valutare se tra due variabili quantitative X e Y esiste un legame lineare e se quest'ultimo è positivo (all'aumentare di una variabile aumenta anche l'altra), negativo (all'aumentare di una variabile, l'altra diminuisce) o neutro (non esiste associazione lineare).

2.1.2 Errore di Tipo *M*, Errore di Tipo *S* e Potenza

Secondo la prospettiva di Neyman & Pearson, nel momento in cui si valutano le ipotesi nulla (H_0) e alternativa, e si prende una decisione a favore di una delle due, possono essere commessi due tipi di errori (Tab.2.1):

- Il primo consiste nell'accettare H_1, quando H_0 è vera. Questo tipo di errore prende il nome di Errore di primo tipo (α)
- Il secondo consiste nel rifiutare H_1, nel momento in cui essa avrebbe dovuto essere accettata. Questo errore prende il nome di Errore di secondo tipo (β).

	H_0 *vera*	H_1 *vera*
$AccettoH_0$	Decisione Corretta (1 – α)	Errore di Tipo II (β)
$AccettoH_1$	Errore di Tipo I (α)	Decisione corretta (1 – β)

Tab.2.1. Possibili decisioni nella prospettiva di Neyman & Pearson: 1 – α rappresenta la probabilità di accettare H_0 quando essa è vera, mentre 1 – β la probabilità di accettare H_1 quando essa è vera (ciò corrisponde alla potenza statistica)

Oltre all'errore di Tipo I e di Tipo II, Gelman e Carlin (2014) hanno formulato altri due tipi di rischi inferenziali, quali l'Errore di Tipo *M* (Magnitudo) e di Tipo *S* (segno).

L'errore di Tipo *M*, definito anche Rapporto di esagerazione, rappresenta la sovrastima media attesa di un effetto rilevato statisticamente significativo. È importante notare che, a differenza degli altri errori, questo non rappresenta una probabilità ma un rapporto, indicando la percentuale media di inflazione dei risultati ottenuti (Bertoldo et al, 2022). Idealmente, l'obiettivo del ricercatore dovrebbe essere quello di ottenere un errore di Tipo *M* il più vicino possibile ad 1, per aumentare le probabilità di trovare una stima dell'effetto il più accurata possibile.

L'errore di Tipo *S* è la probabilità di trovare un risultato statisticamente significativo in direzione opposta a quella plausibile (Bertoldo et al, 2022). Ottenere un Errore di Tipo *S* vicino a 0 indica che i risultati ottenuti vanno in direzione coerente con l'effetto desiderato.

Come si può notare dalla Figura 2.1, questi due rischi inferenziali dipendono dal livello di potenza dello studio, infatti:

- L’errore di Tipo *S* decresce all’aumentare della potenza fino a valori di potenza all’incirca di 0.2, per poi essere approssimabile a 0.0 per valori di potenza superiori.
- L’errore di Tipo *M* è elevato quando la potenza è inferiore a 0.5, ed è via via sempre più approssimabile ad 1. Ciò implica che con potenze superiori vi è un rischio minore di sovrastimare l’effetto ottenuto.

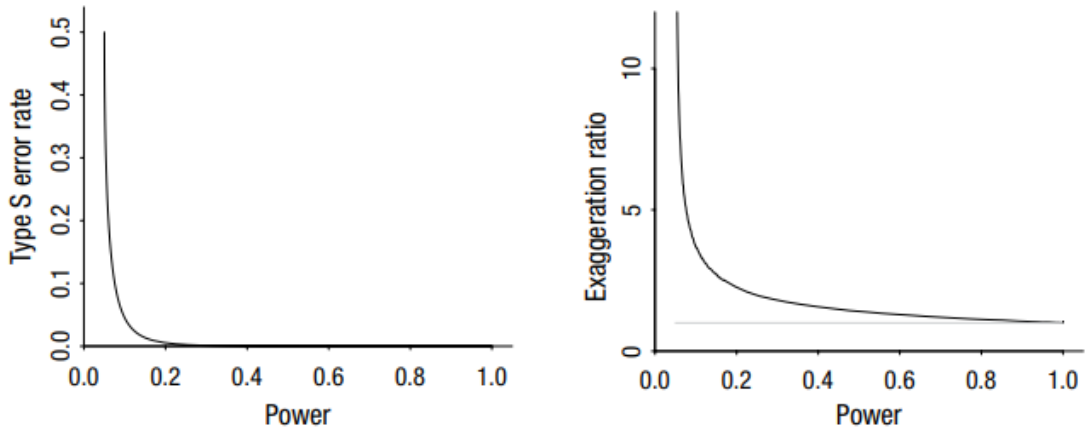

Figura 2.1. Rappresentazione grafica dell’andamento di Errore di Tipo S ed Errore di Tipo M in funzione della potenza (German & Carlin, 2014).

Da ciò consegue che, per ottenere livelli minimi di errore di Tipo *M* ed Errore di Tipo *S* è richiesta una potenza elevata. Livelli maggiori di potenza sono ottenibili aumentando la numerosità campionaria. In figura 2.2, quindi, viene mostrato l’andamento di questi tre indici (Errore di Tipo *M*, Errore di Tipo *S* e potenza) in funzione del campione.

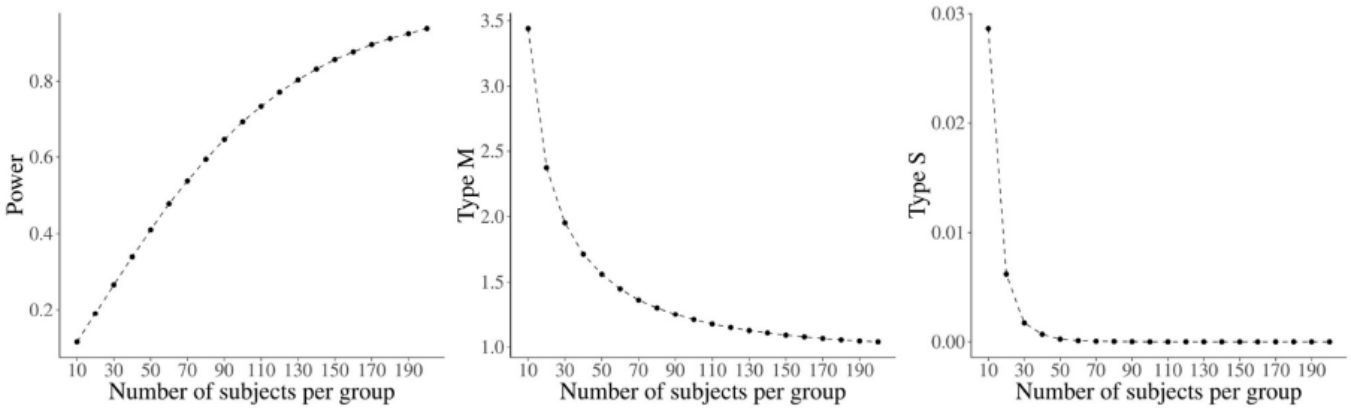

Figura 2.2. relazione tra potenza, Errore di Tipo M, Errore di Tipo S e numerosità campionaria (Altoè et al, 2020),

La Figura 2.2 è tratta da una simulazione di Altoè et al (2020), per cui viene utilizzato un t-test su campioni indipendenti, con una *d* di Cohen di 0.35 e un Errore di Tipo I fissato a 0.05. Si può notare che:

- La potenza cresce all'aumentare del campione e, con una numerosità pari a 130, raggiunge un livello soddisfacente pari a 0.8. L'errore di Tipo *M* diminuisce all'aumentare del campione e, per valori pari o superiori a 190, la sovrastima dell'effetto tende sempre di più ad 1.
- L'errore di Tipo *S* mostra un andamento simile all'Errore di Tipo *M*, divenendo trascurabile per un campione pari o superiore a 60/70 persone.

Come mostrato dalla figura, dunque, valutare una adeguata numerosità campionaria risulta indispensabile per ottenere risultati quanto più possibile attendibili.

2.1.3 Effect size plausibile

Come si è potuto notare dalle definizione degli Errori di tipo *M* e di tipo *S* precedentemente date, essi risultano strettamente legati all'effect size plausibile (Bertoldo et al, 2022). È, quindi, doveroso porre attenzione su di esso e darne una definizione.

L'effect size plausibile è un'assunzione che i ricercatori fanno riguardo all'effetto atteso all'interno della popolazione considerata (Bertoldo et al, 2022). Esso non deve essere tratto dai risultati ottenuti da un singolo studio pilota ma, al contrario, deve essere stabilito a seguito di un'accurata valutazione della letteratura presente, prendendo in considerazione, ad esempio, fonti teoriche, revisioni della letteratura esistente o meta-

analisi, sempre tenendo conto di eventuali bias di pubblicazione (Borenstein et al, 2009, Altoè et al, 2020).

Le fonti teoriche forniscono una base concettuale per formulare ipotesi sull'effect size. Queste fonti includono modelli concettuali, teorie e ipotesi di ricerca sviluppati nell'ambito di ricerca. L'effect size plausibile può derivare da previsioni teoriche sulla relazione tra le variabili coinvolte.

Le revisioni sistematiche della letteratura rappresentano una metodologia per sintetizzare e valutare criticamente studi precedenti su un determinato argomento. Mediante la valutazione dei risultati, esse consentono di ottenere una stima della dimensione dell'effetto basata sulla media dei risultati osservati.

Le meta-analisi, infine, combinano quantitativamente i dati provenienti da diversi studi per ottenere una stima aggregata della dimensione dell'effetto. Questa metodologia statistica consente di integrare e sintetizzare i risultati di studi indipendenti, fornendo una stima più precisa e affidabile dell'effect size.

Un'altra strategia utile per stabilire un effect size plausibile è costituita dalla consultazione di esperti nel settore di interesse (O'Hagan,2018). Essa consente di esprimere le conoscenze di esperti sotto forma di distribuzioni di probabilità. Tale pratica può essere eseguita seguendo due approcci, uno di tipo matematico e l'altro comportamentale e, in entrambi i casi, devono essere rispettati specifici protocolli. Per quanto riguarda l'aggregazione matematica, vengono richieste valutazioni agli esperti e viene calcolata una distribuzione di probabilità per ciascuna distribuzione. Queste vengono poi combinate in un'unica distribuzione aggregata. Per l'aggregazione comportamentale, invece, viene chiesto ad un gruppo di esperti di discutere le loro conoscenze e opinioni fino a raggiungere delle valutazioni di consenso di gruppo, per le quali verrà successivamente calcolata la distribuzione aggregata.

2.2 Design Analysis Prospettiva e Retrospettiva

Come detto precedentemente, la Design Analysis può essere condotta sia nelle fasi preliminari di uno studio, in maniera prospettiva, sia una volta che lo studio è terminato, in maniera retrospettiva. Nel primo caso si tratta di una Design Analysis prospettiva e viene messa in pratica con il fine di identificare una numerosità campionaria sufficiente, tale da massimizzare la potenza e minimizzare gli errori di tipo *M* e di tipo *S*.

Nel secondo caso, invece, si parla di Design Analysis retrospettiva. Essa viene utilizzata per analizzare la potenza dello studio e i diversi errori nel momento in cui sono già stati ottenuti i risultati, consentendone una migliore interpretazione (Bertoldo et al, 2022).

Al fine di dare un'esemplificazione pratica della conduzione di queste due tipologie di DA, nei paragrafi successivi verrà preso in considerazione un case study presentato da Bertoldo et al (2022), che analizza uno studio pubblicato da Eisenberg et al (2003) intitolato "*Does Rejection Hurt? An fMRI Study of Social Exclusion*". L'articolo in questione si basa su un ipotetico coinvolgimento della Corteccia Cingolata Anteriore (ACC) nell'esperienza di dolore causata dal rifiuto sociale. Per testare questa ipotesi, è stato utilizzato un campione costituito da 13 partecipanti, ognuno dei quali doveva giocare ad un videogame insieme ad altri due giocatori (confederati) mentre analizzato tramite risonanza magnetica funzionale (fMRI). I due confederati erano, in realtà, fittizi e il gioco effettivo era svolto dal computer. I giocatori dovevano passarsi la palla in tre diverse condizioni, una di inclusione sociale, una di esclusione sociale esplicita e una di esclusione sociale implicita. Alla fine di ogni esperimento, ogni partecipante doveva compilare un questionario self-report riguardo la propria esperienza di stress durante la fase di esclusione esplicita. Dall'analisi dei dati è risultato un coefficiente di correlazione $r(11) = .88$ e un p-value < 0.05, da cui gli autori hanno concluso che il dolore fisico e quello sociale condividessero gli stessi meccanismi neurali.

2.2.1 Design Analysis Prospettiva

Al fine di stabilire l'effect size plausibile è stata presa in considerazione l'ipotesi di Vul e Pashler (2017), secondo la quale la correlazione tra misure di personalità e attività neurale può essere stabilita di circa $\rho=0.25$. Una correlazione di $\rho=0.50$ è stata considerata plausibile ma troppo ottimista da ottenere e una di $\rho = 0.75$ è stata ritenuta teoricamente plausibile ma irrealizzabile. Inoltre, per considerare entrambe le direzioni di una possibile correlazione, è stato utilizzato nell'analisi un test a due code.

Viene, dunque, condotta una Design Analysis prospettica utilizzando la funzione pro_r(), contenuta in R, che necessita in input di (Figura 2.1):

- Effect size plausible: in questo caso, quindi, 0.25)
- Potenza: ricercata a un livello pari all'80%, quindi $1 – \beta = 0.8$

- Livello di significatività statistica: stabilito a 0.05

```
> pro_r(rho = .25, power = .8, sig_level = .05,
+       alternative = "two.sided", seed = 2020)

        Design Analysis

Hypothesized effect: rho = 0.25

Study characteristics:
   n     alternative   sig_level
   125   two.sided     0.05

Inferential risks:
   power   typeM   typeS
   0.806   1.111   0

Critical value(s): r =  ±0.176
```

Figura2.1. Input e output della funzione pro_r(), utilizzata nel case study preso in considerazione (Bertoldo et al., 2022)

Dai risultati (Tab. 2.2) si evince che per ottenere la potenza prefissata si necessita di un campione costituito da 125 soggetti. Con questa numerosità campionaria, i risultati statisticamente significativi sovrastimerebbero in media l'effetto dell'11% (Errore di Tipo $M = 1.111$) e l'effetto andrebbe in direzione coerente con quello atteso (Errore di Tipo $S = 0$) .

È inoltre emerso un valore critico $r = \pm 0.176$, che consentirebbe, considerati i dati ipotizzati, di confermare l'ipotesi alternativa.

Al fine di investigare come cambierebbero i rischi inferenziali al variare di effect size e potenza, sono state svolte ulteriori simulazioni, riportate in Tabella 2.2.

Si può notare che:

- Diminuendo la potenza desiderata e lasciando invariato l'effect size plausibile, è richiesta una numerosità campionaria di sole 76 persone. In questo modo, però, l'errore di tipo *M* salirebbe al 28% e il valore r critico arriverebbe a ± 0.226.
- Diminuendo l'effect size a 0.15 e lasciando invariata la potenza, invece, si avrebbe bisogno di un campione di 344 persone, nettamente superiore rispetto al primo caso. La sovrastima salirebbe arrivando al 12% e il valore critico diminuirebbe a ± 0.106.

- Infine, aumentando l'effect size a 0.35 e lasciando la potenza a 0.8, la numerosità campionaria necessaria diminuirebbe e il valore r critico aumenterebbe a ± 0.254, portando alla possibilità di accettare l'ipotesi nulla.

In tutti i casi, comunque, l'errore di tipo *S* può essere considerato trascurabile.

ρ	Potenza	N. campionaria	Errore tipo *M*	Errore tipo *S*	*r* critico
0.25	0.8	125	1.111	0	± 0.176
0.25	0.6	76	1.280	0	± 0.226
0.15	0.8	344	1.116	0	± 0.106
0.35	0.8	60	1.115	0	± 0.254

Tab.2.2. simulazione di applicazione della Design Analysis prospettica sull'articolo "Does Rejection Hurt? An fMRI Study of Social Exclusion" (Eisenberg, 2003).
Riadattata da Bertoldo et al (2022)

2.2.2 Design Analysis Retrospettiva

Per condurre una DA retrospettiva occorrono informazioni sul design di ricerca e sul valore dell'effect size plausibile.

Nello studio originale non è stato possibile reperire alcune informazioni, poiché non riportate. Dunque sono state supposte:

- Numerosità campionaria = 13
- $\alpha = 0.05$
- Applicazione di un test a due code

Successivamente, è stata utilizzata la funzione di R retro_r(), i cui input e output sono riportati in Figura2.2

```
> retro_r(rho = .25, n = 13, sig_level = .05,
+         alternative = "two.sided", seed = 2020)

        Design Analysis

Hypothesized effect: rho = 0.25

Study characteristics:
   n    alternative   sig_level
   13   two.sided     0.05

Inferential risks:
   power   typeM   typeS
   0.127   2.583   0.028

Critical value(s): r =  ±0.553
```

Figura 2.2. Input e output della funzione retro_r() applicata al case study preso in considerazione. (Bertoldo et al, 2022)

Svolgendo una Design Analysis retrospettiva su questo articolo, ciò che emerge è che le valutazioni effettuate sembrerebbero erronee.

Per prima cosa, infatti, si può notare che il valore *r* critico è pari a ± 0.55, valore che includerebbe l'effect size plausibile nella regione di accettazione dell'ipotesi nulla. Risulta, pertanto, impossibile trovare contemporaneamente un risultato statisticamente significativo e stimare un effetto vicino a ρ (=0.25).

Lo studio, inoltre, risulterebbe essere incorso in diversi errori inferenziali per quanto riguarda la stima dell'effect size, come evidenziato dai valori dell'Errore di tipo *M* (= 2.582) e dell'Errore di tipo *S* (0.028). Considerando la dimensione di questi errori ne deriva che, qualora venisse identificato un effetto significativo (come in questo caso), esso sovrastimerebbe il valore plausibile di circa due volte e mezzo e ci sarebbe una probabilità di circa il 3% che il risultato statisticamente significativo ottenuto vada in direzione opposta rispetto a quella prevista (in questo caso, quindi, ci sarebbe una probabilità del 3% di avere una relazione negativa tra l'attivazione della ACC e la percezione di dolore sociale).

Infine, come si evince dalla DA effettuata, anche la potenza relativa a questo studio sembrerebbe bassa, approssimabile al 13%. Ciò implica una probabilità del 13% di rifiutare l'ipotesi nulla nel momento in cui esiste un effetto con una correlazione di ρ = |.25|. Sebbene la probabilità di rifiutare H_0 sia bassa, non è impossibile che questo

accada e ciò porterebbe gli sperimentatori a credere che il risultato sia ancora più degno di nota (Gelman & Loken, 2014)

Ciò che si può evincere da questo esempio è, dunque, l'importanza di un'analisi che preceda lo svolgimento di uno studio, in modo tale che esso sia pianificato per essere il più statisticamente accurato.

L'esempio appena discusso è emblematico della cosiddetta euristica della "*Winner's Curse"* (Altoè et al, 2020), per cui, quando la potenza dello studio è bassa, aumenta la probabilità di riscontrare un effetto statisticamente significativo che, però, esageri la dimensione dell'effetto reale (Figura2.3). Ottenere risultati significativi in condizioni di scarsa potenza porta i ricercatori a sovrastimare l'importanza del proprio studio che, nonostante le condizioni svantaggiate di partenza, è riuscito comunque a rilevare un effetto (Bertoldo et al, 2022)

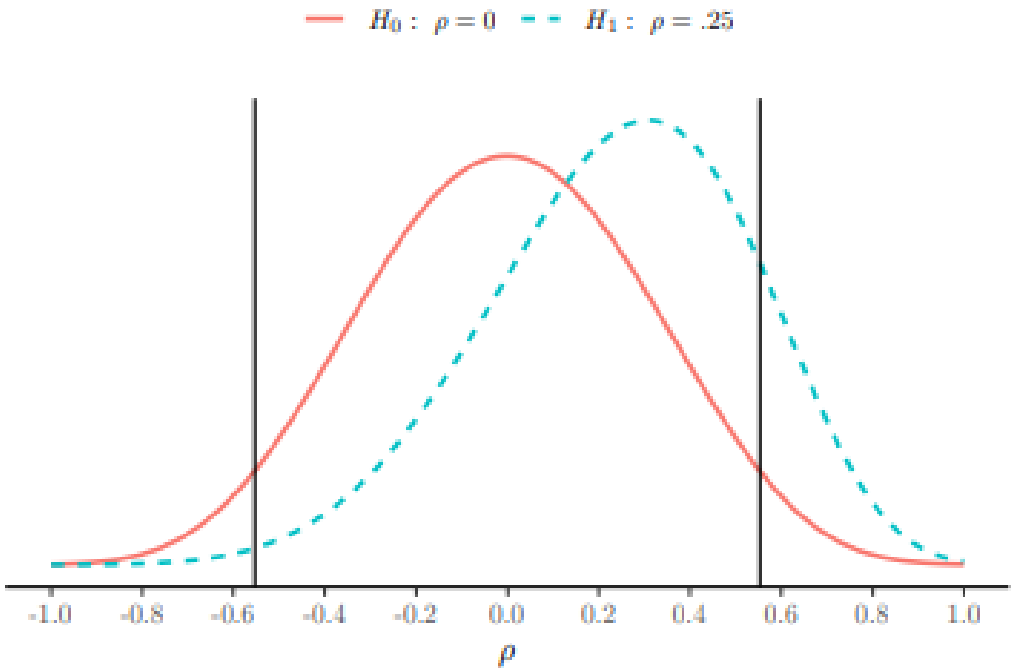

Figura2.3. "Winner's Curse". H_0: Ipotesi nulla, per cui ρ=0; H_1: Ipotesi alternativa, per cui ρ=0.25. Nel caso preso in considerazione, l'effect size plausibile (ρ=0.25) cade nella regione di non rifiuto dell'ipotesi nulla, delimitata dalle linee verticali (r critico= ±0.55). Un risultato ritenuto statisticamente significativo ottenuto da questo studio risulta, per forza di cose, sovrastimare l'effect size plausibile., poiché impossibile riscontrare un risultato statisticamente significativo che sia, allo stesso tempo, vicino all'effetto plausibile. (Bertoldo et al, 2022)

Capitolo 3 - Valutare la differenza tra due coefficienti di correlazione via Design Analysis

Il seguente capitolo rappresenta il cuore di questo elaborato di tesi. Si discuterà, quindi, di come testare la differenza tra due coefficienti di correlazione, concentrandosi soprattutto su differenze di coefficienti provenienti da campioni indipendenti. Verrà, dunque, descritto cosa questo significa e, successivamente, verranno presentate due funzioni di R (*retro_rho2* e *prosp_rho2*) che applicano la Design Analysis al test per la differenza tra correlazioni. Verrà, infine, proposto un approfondimento riguardo la relazione tra potenza e coefficienti di correlazione.

3.1 Valutare la differenza tra due coefficienti di correlazione

Sono numerosi gli studi che si pongono l'obiettivo di determinare la relazione tra due variabili e, per raggiungere tale scopo, viene utilizzato soprattutto il coefficiente di correlazione *r* di Pearson, precedentemente discusso (vedi paragrafo 2.1.1).

Un errore comune quando si effettua un confronto tra due effetti è quello di valutare esclusivamente la differenza tra i livelli di significatività statistica, piuttosto che il livello di significatività della differenza tra i due (Nieuwenhuis et al, 2011). Si consideri, ad esempio, uno scenario estremo in cui viene valutata l'efficacia di un farmaco, somministrato ad un campione sperimentale, rispetto ad un placebo, somministrato ad un campione di controllo. Si ponga caso che l'effetto rilevato per il campione di controllo sia considerato statisticamente significativo (p = 0.048), al contrario dell'effetto riscontrato nel campione di controllo (p = 0.053). Nonostante il fatto che questi due valori siano opposti rispetto alla soglia di significatività del 5%, non è possibile concludere che l'effetto del farmaco differisca statisticamente da quello del placebo. Rosnow & Rosental (1989) affermano, a tal riguardo, che "Dio ama il 0.06 quasi quanto il 0.05"[2], proprio ad

[2] "Surely, God loves the 0.06 nearly as much as the 0.05" Rosnow, R.J. & Rosenthal, R. Statistical procedures and the justification of knowledge in psychological science. Am. Psychol. 44, 1276–1284 (1989)

indicare che differenze così minime non implicano direttamente una differenza negli effetti.

Quando si effettua un confronto tra due correlazioni i ricercatori non dovrebbero, quindi, soffermarsi eslusivamente sul livello di significatività, ma dovrebbero valutare soprattutto la significatività statistica della loro differenza (Gelman & Stern, 2014; Nieuwenhuis et al, 2011).

Bisogna, quindi, considerare in che contesto effettuare tale valutazione. Gli scenari possibili sono due (Diedenhofen et al, 2015) :

1. Le correlazioni misurate sono state ottenute da due campioni indipendenti A e B ($H|0: \rho_A - \rho_B = 0$). Es: testare la correlazione tra empatia e bullismo in maschi e femmine.
2. Le correlazioni tra le variabili (1,2,3,4) provengono dallo stesso campione.
   Vanno, dunque, differenziati due casi:
   a) Le due correlazioni hanno una variabile in comune ($H_{0:}\rho_{12} - \rho_{23} = 0$).
      Es: testare se la correlazione tra empatia e bullismo sia diversa rispetto alla correlazione tra bullismo e status socioeconomico nello stesso campione di soggetti
   b) Le due correlazioni non sono sovrapponibili ($H_0: \rho_{12} - \rho_{34} = 0$
   c) Es: testare se la correlazione tra empatia e bullismo sia diversa rispetto alla correlazione tra status socioeconomico e creatività, all'interno dello stesso campione.

Questo elaborato di tesi verterà esclusivamente sul primo caso citato, ovvero il test per la differenza di coefficienti di correlazione provenienti da due campioni indipendenti. Nello specifico, utilizzando una procedura basata sulla simulazione.

3.2 Funzioni in R per la design analysis

Per testare la differenza tra due coefficienti di correlazione su campioni indipendenti, utilizzando la Design Analysis, sono state costruite due funzioni per il software statistico R, quali *retro_rho2* (per la retrospective DA) e *prosp_rho2* (per la prospective DA), consultabili in appendice (Appendice A). Affinché esse funzionino necessitano di due pacchetti di R, quali *cocor* e *mass*, entrambi scaricabili dal sito di R.

Entrambe le funzioni sono basate su un processo di simulazione di Monte Carlo, analogo a quello adottato da Bertoldo et. al 2022, adattato al caso della differenza tra coefficienti di correlazione.
Per procedere con l'utilizzo delle funzioni sarà sufficiente copiare il testo delle funzioni riportato in Appendice e mandarlo in esecuzione nella Console di R.

3.2.1 Introduzione all'utilizzo delle funzioni

Scopo delle due funzioni presentate è quello di valutare la differenza di coefficienti di correlazione provenienti da due campioni indipendenti, tramite test a due code e utilizzando la Prospective e la Retrospective Design Analysis.
Per farlo bisogna, innanzitutto, ricavare la differenza di correlazione plausibile, cercando gli indici necessari tramite revisione della letteratura, meta-analisi o elicitazione degli esperti. Solo una volta ottenuti i coefficienti e svolta la loro differenza sarà possibile applicare la Design Analysis.

Nel caso di Prospective DA, la differenza di coefficienti di correlazione plausibile sarà necessaria per ricavare la numerosità campionaria adeguata per ottenere un livello di potenza sufficientemente buono (fissato nella funzione *prosp_rho2* a .80 ma facilmente modificabile dall'utente); nel caso della Retrospective DA, invece, essa verrà utilizzata, assieme alla numerosità campionaria dello studio considerato, per stabilire se tale studio, per quella determinata differenza di correlazioni plausibile, risulti sottopotenziato o meno.

Per entrambe le funzioni è stato necessario effettuare un doppio passaggio: in input vengono indicati i coefficienti di correlazione plausibili ipotizzati e successivamente viene effettuato il calcolo per la differenza tra tali indici. Ciò è dovuto al fatto che per una stessa differenza di correlazione plausibile, al variare degli indici coinvolti in tale differenza, si ottengono potenze differenti e sarebbe stato, dunque, insufficiente inserire esclusivamente la differenza di correlazione da valutare. Una dimostrazione pratica della relazione tra potenza e differenze di coefficienti di correlazione sarà presentata nel Paragrafo 3.3.

3.2.2 *retro_rho2*

La funzione *retro_rho2* (Figura 3.1) serve per effettuare una Retrospective Design Analysis sui risultati di uno studio.

```
> retro_rho2(rho1,n1,rho2,n2,B,alpha)
```

Figura 3.1. Formulazione di *retro_rho2*

Come mostrato in Figura 3.1, la funzione necessita in input di:

1. **rho1**: coefficiente di correlazione plausibile nella prima popolazione presa in considerazione;
2. **n1**: numerosità campionaria del primo campione;
3. **rho2**: coefficiente di correlazione plausibile nella seconda popolazione presa in considerazione
4. **n2**: numerosità campionaria del secondo campione. Nel caso in cui essa non venga espressa, viene di default considerata pari a n1;
5. B: numero di repliche della simulazione. È impostato di default a 1000, ma può essere facilmente aumentato per ottenere risultati maggiormente stabili
6. alpha: errore di primo tipo fissato a 0.05 di default, ma che può essere modificato dall'utente a seconda delle esigenze

e restituisce in output:

- potenza
- Errore di tipo M
- Errore di tipo S

Per procedere con la funzione è sufficiente indicare gli indici riportati in grassetto (rho1, n1, rho2, n2), facendo attenzione che rho1 > rho2. In caso contrario, R segnalerà il seguente errore (Figura 3.2):

```
> retro_rho2(.1,750,.3,750)
Error in retro_rho2(0.1, 750, 0.3, 750) :
  First correlation must be greater than second correlation
```

Figura3.2. Errore segnalato da R nel momento in cui rho1 < rho2. Per risolverlo sarà sufficiente indicare i due coefficienti in ordine decrescente.

Per una migliore comprensione della funzione, di seguito sarà riportato un esempio di applicazione.

Supponendo che un gruppo di ricercatori abbia voluto testare la correlazione tra autolesionismo non suicidario (NSSI) e sentimenti di colpa su due campioni indipendenti costituiti da 750 maschi e 750 femmine, l'analisi statistica effettuata rileva un coefficiente di correlazione pari a r_1 = .10 per i maschi e r_2 = .20 per le femmine, riscontrando, quindi, una differenza di correlazione pari a 0.1, che viene considerata significativa. A fini illustrativi e in assenza di ulteriori dati, si ponga caso che i coefficienti di correlazione plausibili ricavati dalla letteratura siano per la popolazione maschile ρ_1 = 0.20 e per quella femminile ρ_2 = 0.30 (differenza di correlazioni plausibile = 0.1). Utilizzando tali dati, viene in seguito effettuata una Retrospective Design Analysis (Figura 3.3)

```
> retro_rho2(.3,750,.2,750)
$power
[1] 0.511

$typeS
[1] 0

$typeM
[1] 1.316759
```

Figura 3.3. Applicazione della funzione Retro_rho2

Come si può notare dalla Figura 3.3, vengono forniti in input i due coefficienti di correlazione e le numerosità campionarie. In output, vengono restituiti potenza, Errore di Tipo S ed Errore di Tipo M.

Dai risultati ottenuti dall'applicazione della Retrospective DA si può desumere che lo studio in questione sia sottopotenziato per valutare una differenza tra coefficienti di correlazione plausibile pari a 0., rilevando una potenza di 0.5 e l'Errore di tipo *M* del 31%.

3.2.3 *prosp_rho2*

L'applicazione della funzione *prosp_rho2* (Figura 3.4), invece, è utile per calibrare la numerosità campionaria di uno studio rispetto ad una potenza ed Errori di Tipo *M* ed *S* desiderati.

```
> prosp_rho2(rho1,rho2,alpha,power,rangen,tol,iter,verbose)
```

Figura 3.4. Formulazione di *prosp_rho2*

La funzione richiede, dunque, in input:

1. **rho1**: effect size plausibile riferito alla prima popolazione
2. **rho2**: effect size plausibile riferito alla seconda popolazione
3. alpha: errore di primo tipo, fissato a 0.05 per default e modificabile dall'utente
4. power: potenza, fissata a 0.80 per default e modificabile dall'utente
5. B: numero di repliche della simulazione. È impostato di default a 1000(=1000) ma può essere facilmente aumentato per ottenere risultati maggiormente stabili
6. rangen: range all'interno del quale viene ricercata la numerosità campionaria adeguata a raggiungere la potenza desiderata . Di default è fissato tra 10 e 2000 ma può essere modificato, come avviene nell'esempio di seguito riportato
7. tol: tolleranza nella ricerca della potenza. In particolare, la funzione considera come potenza ottimale nel calcolo della numerosità campionaria una potenza che stia nell'intervallo tra potenza desiderata - tolleranza e potenza desiderata + tolleranza. Di default la tolleranza è fissata all'1%, può essere modificata dall'utente
8. iter: numero massimo di volte in cui viene ripetuto il ciclo per la ricerca della potenza
9. verbose: un valore corrispondente a "TRUE" (o "T") produce in output non solo i risultati finali, ma anche tutti i passaggi intermedi svolti dalla funzione. Il default è "FALSE" (o "F") e implica che solo i risultati finali saranno prodotti in output.

E fornisce:

- Potenza stimata
- Numerosità campionaria necessaria
- Errore di Tipo M

- Errore di Tipo S

Anche con *prosp_rho2* sarà sufficiente indicare i valori riportati in grassetto (rho1, rho2),

Supponendo di trovarsi nel caso citato nel Paragrafo 3.2.2 e ipotizzando, dunque, di avere una differenza di correlazioni plausibile di 0.1 (ρ_1 = 0.20, ρ_2 = 0.30), e la correlazione dei due gruppi pari a r_1 = .10 e r_2 = .20, sarà di seguito applicata la Design Analysis prospettica (Figura 3.5).

```
> prosp_rho2(.3,.2)
$power
[1] 0.8

$est_power
[1] 0.794

$n_group
[1] 1333

$typeM
[1] 1.121832

$typeS
[1] 0
```

Figura 3.5. Applicazione di *prosp_rho2* ad un caso ipotetico.

Dall'applicazione si evince che, se si dovesse pianificare la numerosità campionaria dello studio in questione, per una differenza plausibile di 0.1, dati i due coefficienti considerati, si avrà bisogno di una numerosità campionaria costituita da 1333 soggetti. Tale numerosità consente di ottenere una potenza di 0.79, un Errore di Tipo *S* nullo ed una sovrastima del 12%.

3.3 Approfondimento: relazione tra potenza e differenza di correlazioni

In questo paragrafo verrà presentata e discussa un'applicazione della funzione *retro_rho2* ad un caso ideale, a dimostrazione della relazione che intercorre tra differenza di correlazione e potenza.

Considerata una numerosità campionaria fissata a 500, è stata valutata la relazione tra potenza e differenza di correlazioni per campioni indipendenti, tramite l'utilizzo della funzione *retro_rho2*(Figura 3.6), di seguito riportata:

```
> retro_rho2(rho1,500,rho2,500,B= 10000)
```

Figura 3.6. Applicazione di *retro_rho2* al caso preso in analisi. Al variare di ρ_1 e ρ_2, la loro differenza ($\Delta\rho$ è tenuta costante a 0.1

Sono stati presi in considerazione i coefficienti di correlazione tra ρ = -1 e ρ = +1, tenendo costante la differenza di correlazioni ($\rho_1 - \rho_2 = 0.1$). Il punto che sull'asse delle Y corrisponde ad un valore sull'asse delle X di – 0.9 corrisponde alla potenza associata alla differenza tra i coefficienti di correlazione – 1 e – 0.9. Sono stati considerati tutti i coefficienti tra -1 e +1 per un totale finale di 20 punti complessivi.

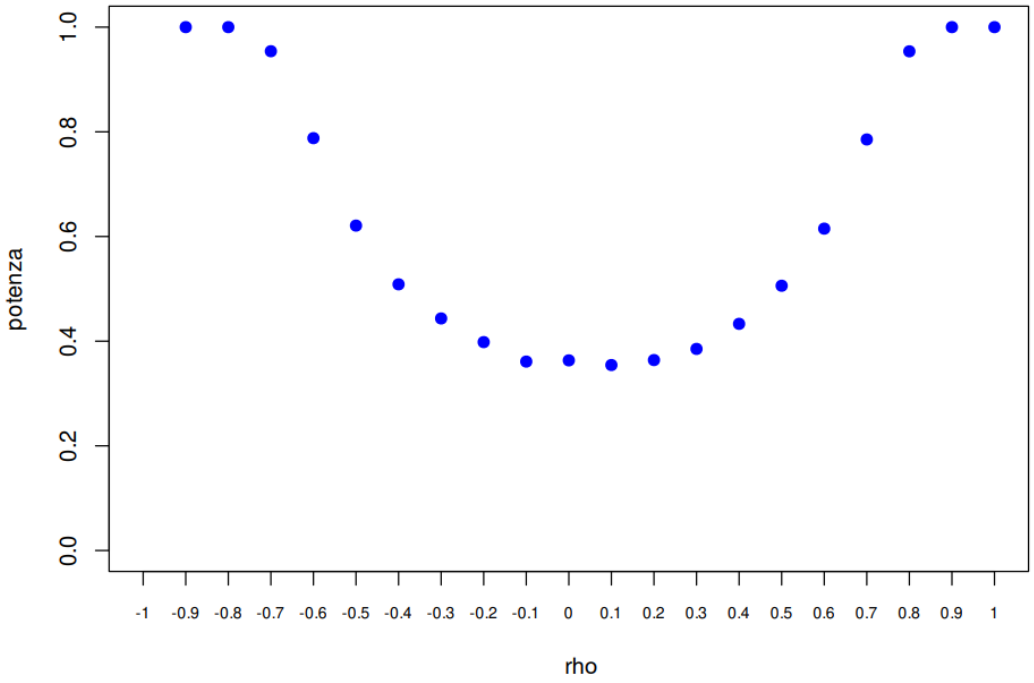

Figura 3.7. Relazione tra potenza e differenza tra correlazioni per campioni indipendenti pari a .1 al variare dei valori dei due indici di correlazione coinvolti nella differenza. Ad esempio, il valore sull'asse delle *Y* che corrisponde a un valore sull'asse delle *X* di .3, è la potenza associata ad una differenza tra i coefficienti di correlazione .3 e .2. La numerosità campionaria è stata fissata a 500 per ciascun campione indipendente.

Dal grafico in Figura 3.7 si può evincere che la relazione che intercorre tra la potenza e le differenze tra coefficienti di correlazione può essere rappresentata come una curva ad "U": all'aumentare dei valori assoluti dei ρ coinvolti nella differenza, la potenza ad essi

associata aumenta. Ciò dimostra che più che la differenza stessa, in questo caso fissata a 0.1, è rilevante il punto da cui questa differenza parte, per cui, ad esempio, la potenza associata alla differenza tra i coefficienti $\rho_1 = 0.9$e $\rho_2 = 0.8$ è superiore rispetto a quella associata alla differenza tra $\rho_3 = 0.3 e \rho_4 = 0.2$.

Tenendo in considerazione gli Errori di Tipo *M* e di Tipo *S*, si può inoltre dedurre che essi, fissata una differenza tra coefficienti di correlazione e fissate le rispettive numerosità dei due campioni, diminuiscono all'aumentare in valore assoluto dei coefficienti di correlazione coinvolti nel calcolo della differenza prefissata, mostrando un andamento opposto rispetto a quello della potenza. Ad esempio, considerando la differenza tra $\rho_1 e \rho_2$, essa sarà associata ad un Errore di Tipo *M* nettamente inferiore rispetto a quello associato alla differenza tra $\rho_3 e \rho_4$.

Capitolo 4 – Case study: Applicazione della Design Analysis ad un caso reale

4.1 Lo studio originale

Nel 2009 la rivista "*Journal of Research in Reading*" ha pubblicato l'articolo "*Gender differences in reading ability and attitudes: examining where these differences lie*" (Logan & Johnston, 2009), il cui obiettivo era investigare le differenze di genere nella relazione tra abilità di lettura, frequenza di lettura e atteggiamenti e credenze legate alla lettura e alla scuola.

Nello studio è stato considerato un campione costituito da 232 soggetti madrelingua inglese (117 maschi, 115 femmine), di età compresa tra i 10 anni e gli 11 anni e 9 mesi (età media 10 anni e 7 mesi, ds = 0.35), provenienti da otto scuole primarie differenti, situate in aree con status socioeconomico variabile da basso ad alto. Tutti i bambini sono stati testati tra la fine del sesto anno di scuola e l'inizio del settimo.

I bambini hanno dovuto compilare un test di abilità di lettura, legato alla comprensione del testo, e un questionario. Nello specifico:

- Il *Reading ability. Group Reading Test II*, test composto da 45 domande, che misurano la lettura di parole, la comprensione e il vocabolario.
- Un questionario self-report progettato dai ricercatori, composto da 14 domande valutate con scala Likert a 5 punti e costituito da ulteriori due domande introduttive sulla frequenza di lettura e su quanto spesso i bambini prendessero libri dalla biblioteca. Lo strumento è stato utilizzato per ottenere informazioni sull'atteggiamento verso la lettura (5 domande) e verso la scuola (5 domande), sulle credenze riguardo la propria competenza (2 domande) e sul supporto accademico percepito, rispetto a pari e insegnanti (2 domande).

Lo studio ha rilevato che le bambine, in generale, dimostrano una migliore comprensione della lettura, una maggiore frequenza e manifestano un atteggiamento più positivo verso la lettura e verso la scuola rispetto ai ragazzi. Sono state individuate ulteriori differenze di genere (Tabella 4.1) per quanto riguarda la relazione tra i fattori valutati dal questionario. I bambini hanno, infatti, mostrato correlazioni significativamente più forti delle bambine tra l'atteggiamento positivo verso la scuola e il supporto accademico

percepito (*p-value* < 0.01). Anche la relazione tra atteggiamento verso la lettura e credenze di competenza è risultata essere più forte nei maschi rispetto che nelle femmine, con un *p-value* < 0.01.

	Boys				Girls			
	ATR	ATS	CB	SUP	ATR	ATS	CB	SUP
ATR	–	.33**	.37**	.07	–	.43**	.28**	– .01
ATS	.37**	–	.47**	.27**	.43**	–	.25*	– .07
CB	.42**	.50**	–	.15	.27**	.24*	–	.04
SUP	.07	.27**	.14	–	– .00	– .07	.06	–

Tabella 4.1. Correlazioni tra aree del questionario per maschi e femmine.

Nota: ATR = *attitude to reading,* atteggiamento verso la lettura; ATS = *attitude to school,* atteggiamento verso la scuola, CB = *competency* beliefs, credenze sulla propria competenza, SUP = *support,* supporto percepito dei pari e degli insegnanti. Da Logan et al, 2009.

Il dato su cui si concentrerà l'analisi di questo elaborato di tesi è quello che riguarda la differenza di correlazione tra l'atteggiamento positivo verso la scuola e le credenze sulla propria competenza, ritenuta dall'articolo statisticamente significativa. Lo studio ha, infatti, individuato una differenza di correlazione pari a 0.26, ottenuta da una correlazione di 0.50 nei maschi e 0.24 nelle femmine. Questa differenza verrà di seguito valutata tramite Prospective e Retrospective Design Analysis.

4.2 Applicazione della Design Analysis al caso

Di seguito verranno riportate le applicazioni delle funzioni *prosp_rho2* e *retro_rho2* allo studio precedentemente spiegato.

Per una maggior completezza, lo studio verrà valutato rispetto ad una differenza di coefficienti di correlazione di 0.2 e di 0.1[3]. Come indici plausibili sono state prese in considerazione una correlazione di 0.3 per quanto riguarda i bambini, considerata da Cohen (1988) l'effect size medio in psicologia in assenza di altre informazioni e 0.2 e 0.1 per quanto concerne le bambine. Un'ulteriore valutazione è stata successivamente effettuata mantenendo costante la differenza di 0.2, considerando un coefficiente di correlazione plausibile per i maschi di 0.4 e, di conseguenza, di 0.2 per le femmine.

[3]Considerando che l'approccio statistico presentato in questo lavoro si basa su un test a due code, di fatto valutare una differenza plausibile di .2 corrisponde a valutare come differenze plausibili sia .2 che -.2.

Gli indici sono stati scelti a puro titolo illustrativo e metodologico e non derivano, quindi, da una valutazione della letteratura. Per una procedura più accurata è necessario consultare fonti teoriche, revisioni della letteratura esistente o meta-analisi.

4.2.1 Prospective Design Analysis

Si supponga di essere nella fase di progettazione dello studio in questione e di voler investigare la correlazione tra atteggiamento verso la scuola (ATS) e convinzioni rispetto alle proprie competenze (CB). Si vogliono, quindi, ottenere indicazioni rispetto alla numerosità campionaria necessaria per ottenere una potenza dell'80% per valutare una differenza di coefficienti di correlazione plausibile di 0.1, prendendo come coefficienti di partenza, e quindi effect-size plausibili, 0.3 per i maschi e 0.2 per le femmine. In Figura 4.1 è, dunque, riportata l'applicazione della funzione al caso considerato.

```
> prosp_rho2(.3,.2)
$power
[1] 0.8

$est_power
[1] 0.802

$n_group
[1] 1379

$typeM
[1] 1.128662

$typeS
[1] 0
```

Figura 4.1. Applicazione della Prospective Design Analysis ad una differenza di coefficienti di correlazione di 0.1

L'analisi effettuata rileva che per ottenere una potenza dell'80% sia necessaria una numerosità campionaria di 2758 soggetti (1379 soggetti per gruppo). Con questo campione, l'Errore di Tipo *S* è minimo ed approssimato allo 0, mentre l'Errore di Tipo *M* indica una sovrastima media del 13% (typeM = 1.129).

Per un'analisi più completa, la stessa procedura verrà svolta anche considerando come secondo coefficiente di correlazione 0.1 e, quindi, una differenza di correlazione plausibile di 0.2 (Figura 4.2)

```
> prosp_rho2(.3,.1)
$power
[1] 0.8

$est_power
[1] 0.792

$n_group
[1] 353

$typeM
[1] 1.1374

$typeS
[1] 0
```

Figura 4.2. Applicazione della Prospective Design Analysis ad una differenza di coefficienti di correlazione plausibile di 0.2

In questo caso, essendo la differenza tra coefficienti più elevata, per ottenere la medesima potenza sarà necessario un campione di soggetti nettamente minore, pari a 706 (353 soggetti per gruppo). In queste condizioni si incorrerebbe in una sovrastima media del 14% (typeM = 1.137) e l'Errore di Tipo *S* risulterebbe il medesimo.

Considerando quanto detto nel Paragrafo 3.3 riguardo alla relazione che intercorre tra potenza e differenza di coefficienti di correlazione, è interessante porre attenzione anche a cosa succede nel caso in cui la differenza valutata parta da un altro punto.

Si ponga, quindi, l'ipotesi ulteriore di voler effettuare la Prospective Design Analysis su una differenza di coefficienti di correlazione pari a 0.2, ma utilizzando, questa volta, $\rho_1 = 0.4 e \rho_2 = 0.2$ (Figura 4.3)

```
> prosp_rho2(.4,.2)
$power
[1] 0.8

$est_power
[1] 0.806

$n_group
[1] 326

$typeM
[1] 1.120382

$typeS
[1] 0
```

Figura 4.3. Applicazione della Prospective Design Analysis ad una differenza di correlazione plausibile di 0.2, prendendo come coefficienti 0.4 e 0.2

Come mostrato in Figura 4.3, per ottenere una potenza dell'80%, utilizzando come effect-size plausibile una differenza di 0.2, derivante dalla differenza tra 0.4 e 0.2, saranno necessari due campioni costituiti da 326 persone ciascuno

Dalle analisi effettuate si può ipotizzare, quindi, che la numerosità campionaria utilizzata per lo studio preso in causa risulta, per una differenza di coefficienti di correlazione pari a 0.2 e di 0.1, decisamente inferiore rispetto a quella necessaria per ottenere una potenza adeguata, stimata utilizzando come effect-size plausibili sia 0.4 che 0.3.

4.2.2 Retrospective Design Analysis

Per effettuare una Retrospective Design Analysis sul caso analizzato, bisogna partire dalle informazioni riguardo la numerosità campionaria. In questo studio, il campione di bambini è composto da 117 soggetti e quello formato da bambine da 115 soggetti.

Per quanto riguarda l'effect size plausibile e, quindi, la differenza di coefficienti di correlazione, verranno utilizzati i dati usati per la Prospective DA. Si analizzerà, dunque, una differenza di 0.1 ($\rho_1 = 0.3; \rho_2 = 0.2$, riportata in Figura 4.4 e di 0.2, sia partendo da una correlazione media ($\rho_1 = 0.3; \rho_2 = 0.1$, come mostrato in figura 4.5, che da una correlazione medio-alta ($\rho|1 = 0.4; \rho_2 = 0.2$), in Figura 4.6.

```
> retro_rho2(.3,117,.2,115)
$power
[1] 0.132

$typeS
[1] 0.03787879

$typeM
[1] 3.103105
```

Figura4.4. Applicazione della Retrospective Design Analysis ad una differenza di coefficienti di correlazione di 0.1.

Per la valutazione di una differenza di correlazioni plausibile di 0.2 lo studio si dimostra notevolmente sottopotenziato, ottenendo una potenza di appena di .13. Ciò sta a significare che c'è una probabilità del 13% di rifiutare l'ipotesi nulla, nel caso in cui esistesse una differenza di coefficienti di correlazione di $\Delta\rho_{12} = 0.1$. È, inoltre, evidente che lo studio in questione presenti degli importanti rischi inferenziali per quanto riguarda la stima dell'effect size, come è dimostrato dall'Errore di Tipo *S* (typeS = 0.038) e l'Errore di Tipo *M* (typeM = 3.103). Ciò sta a indicare che, nel caso venisse riscontrato una differenza di coefficienti di correlazione statisticamente significativa, essa risulterebbe fortemente sovrastimata e con una probabilità circa del 4% di andare in direzione opposta a quella dell'effect size plausibile.

In conclusione, il campione utilizzato risulta insufficiente per ottenere una buona potenza.

Di seguito (Figura 4.5) è stato poi analizzato il caso di una differenza di coefficienti di correlazione pari a $\Delta\rho_{12} = 0.2$.

```
> retro_rho2(.3,117,.1,115)
$power
[1] 0.367

$typeS
[1] 0.002724796

$typeM
[1] 1.642163
```

Figura 4.5 Applicazione della Retrospective Design Analysis ad una differenza di coefficienti di correlazione di 0.2

Pur aumentando la differenza tra coefficienti, i risultati ottenuti hanno un andamento simile a quello del caso precedentemente descritto. Emerge, infatti, una potenza di 0.37, una sovrastima del 64% e un Errore di Tipo *S* di circa 0.003.

Se la differenza considerata partisse da un coefficiente di correlazione medio-alto ($\rho_1 = 0.4$), con le numerosità campionarie utilizzate dallo studio, si otterrebbero comunque una potenza e un Errore di tipo *S* paragonabili a quelli del caso precedentemente descritto e un Errore di tipo *M* lievemente inferiore (Figura 4.6).

```
> retro_rho2(.4,117,.2,115)
$power
[1] 0.364

$typeS
[1] 0.002747253

$typeM
[1] 1.574614
```

Figura 4.6. Applicazione della Retrospective Design Analysis alla differenza di coefficienti di correlazione ottenuta da $\rho_1 0.4 e \rho_2 = 0.2$.

In conclusione, i risultati ottenuti mostrano che, almeno per le dimensione dell'effetto considerate ($\Delta\rho_{12} = 0.2 e \Delta\rho_{12} = 0.1$) lo studio è sottopotenziato e, probabilmente, la differenza ottenuta, e considerata statisticamente significativa, è una sovrastima dell'effetto reale.

Dall'analisi effettuata si può pensare di ritrovarsi nel caso della Winner's Curse, dato che la significatività statistica ottenuta sembra derivare da una scarsa potenza e conseguentemente l'effect size stimato nello studio è plausibilmente una sovrastima del vero effect size.

Per ricavare risultati più informativi bisognerebbe, comunque, procedere con la Design Analysis solo a seguito di uno studio dettagliato della letteratura, che consenta di estrapolare gli effect size plausibili per il fenomeno studiato.

Capitolo 5 – Conclusioni

La crisi di replicabilità emersa in questi ultimi decenni potrebbe, potenzialmente, aver segnato un punto di svolta per la metodologia di ricerca e la diffusione dei risultati. Essa ha, difatti, portato i ricercatori ad interrogarsi su come poter invertire la tendenza e rendere i propri lavori di ricerca il più accurati possibile. È in questo periodo che iniziano a diffondersi, ad esempio, piattaforme open source (come Open Science Framework) su cui caricare e, di conseguenza, condividere con la comunità scientifica, tutti i dettagli e le analisi relative ad un determinato lavoro di ricerca (Bertoldo, 2019). Da ciò consegue, dunque, una facilitazione rispetto alle procedure per la replicazione degli studi.

La crisi ha, inoltre, dato l'inizio alla presa di coscienza del fatto che la significatività statistica dovuta esclusivamente all'analisi del *p-value* risulta, tutto sommato, insufficiente. Da qui deriva la necessità di prendere in considerazione anche altre componenti caratteristiche di uno studio come, ad esempio, la potenza e la dimensione dell'effetto. Studi con una bassa potenza possono, infatti, portare a risultati valutati statisticamente significativi ma che risultano fuorvianti e, in alcuni casi, sovrastime eccessive degli effetti plausibili.

È per rispondere a questo bisogno di analisi più approfondite e accurate che entra in gioco la Design Analysis che, tramite la stima di un effect size plausibile, ottenuto attraverso l'analisi della letteratura presente, e la definizione dell'errore di segno e di magnitudo, consente una valutazione migliore dello studio, sia nelle fasi preliminari (Prospective Design Analysis) sia a lavoro compiuto (Retrospective Design Analysis).

Scopo del presente elaborato di tesi è stato, dunque, quello di dare un minimo contributo verso la risoluzione della Crisi, tramite la presentazione della Design Analysis, nello specifico descrivendone le possibili applicazioni per quanto riguarda la valutazione della differenza di due coefficienti di correlazione estratti da campioni indipendenti. Nel contesto della ricerca in psicologia, sono numerosi gli studi in cui ci si prefigge di valutare e confrontare due correlazione in campioni o condizioni differenti. La pratica più comunemente diffusa prevede di confrontare i coefficienti tramite la valutazione dei livelli di significatività statistica. In un'ottica di confronto, però, basarsi esclusivamente

su questo dato risulta insufficiente (Nieuwenhuis et al, 2011) ed è, dunque, necessario concentrarsi anche sulla loro differenza.

È a tale scopo che sono state formulate le due funzioni di R, *prosp_rho2* e *retro_rho2*, presentate nel terzo capitolo, utilizzabili rispettivamente per la prospective e la retrospective Design Analysis. Tramite questi strumenti, definendo in input la differenza di coefficienti di correlazione plausibile posta in analisi, è possibile ottenere informazioni riguardo la numerosità campionaria, la potenza relativa allo studio, l'Errore di Tipo *M*, che rappresenta la sovrastima media attesa dell'effetto, e l'Errore di Tipo *S,* ovvero la probabilità che l'effetto ottenuto vada in direzione opposta rispetto a quello atteso. Considerare anche questi dati consente di migliorare la qualità delle inferenze statistiche, riducendo il rischio di giungere a conclusioni poco plausibili.

Nel terzo capitolo è stata, inoltre, messa in evidenza la relazione che intercorre tra la differenza tra coefficienti di correlazione e la potenza: tramite la funzione *retro_rho2*, infatti, si è potuto constatare come, a parità di differenza di correlazione, la potenza aumenti all'aumentare del valore assoluto dei coefficienti di correlazione coinvolti nella valutazione. Dunque, più le correlazioni prese in analisi sono forti, maggiore sarà la potenza.

I limiti rilevanti riscontrati nel presente elaborato di tesi sono principalmente due. È importante notare, infatti, che le funzioni descritte svolgono test statistici basati su un approccio a due code. Questo significa che vengono valutate esclusivamente le differenze significative tra le variabili in esame, senza considerare specificamente l'eventuale direzione delle correlazioni, non consentendo di catturare completamente la complessità della relazione tra le correlazioni prese in esame. Potrebbe, dunque, essere utile, per raggiungere un livello superiore di completezza e chiarezza, implementare *prosp_rho2* e *retro_rho2* in modo tale che sia possibile ottenere maggiori informazioni riguardo le differenze poste in analisi al fine di valutare, oltre alla significatività delle differenze considerate, anche la direzione in cui vanno tali differenze.

Inoltre, un secondo aspetto rilevante è che le funzioni sono state scritte per valutare solo correlazioni provenienti da campioni indipendenti. Non è stato, dunque, tenuto conto delle correlazioni che potrebbero emergere da campioni dipendenti. Sarebbe interessante, quindi, un ulteriore sviluppo delle funzioni in questa direzione, consentendo di ampliare le possibilità di valutazione anche agli studi in cui le correlazioni provengono

da campioni indipendenti, dato che ricerche di questo tipo sono molto presenti in letteratura.

In conclusione, il principale vantaggio di utilizzare la Design Analysis è quello di mettere in luce aspetti della ricerca che, fino a prima della Crisi di Replicabilità, non sono stati considerati di frequente, come la potenza dello studio, o nuovi rischi inferenziali necessari per ottenere una valutazione più completa dei risultati, come gli Errori di Tipo *M* e di Tipo *S.* Considerata l'importanza di questi fattori, risulta necessario, infine, prendere le distanze dalla comune concezione di significatività statistica fondata esclusivamente sulla valutazione del *p-value*, tipica del Null Hypothesis Significance Testing, a favore di altri metodi che consentano una valutazione più accurata degli effetti e, più in generale, dello studio.

BIBLIOGRAFIA

Alister, M., Vickers-Jones, R., Sewell, D. K., & Ballard, T. (2021). How do we choose our giants? Perceptions of replicability in psychological science. *Advances in Methods and Practices in Psychological Science*, 4(2), 25152459211018199.

Allen, M. S., Iliescu, D., & Greiff, S. (2023). Direct Replication in Psychological Assessment Research. *European Journal of Psychological Assessment*.

Altoè, G., Bertoldo, G., Zandonella Callegher, C., Toffalini, E., Calcagnì, A., Finos, L., & Pastore, M. (2020). Enhancing statistical inference in psychological research via prospective and retrospective design analysis. *Frontiers in Psychology*, 10, 2893.

Banks, G. C., O'Boyle Jr, E. H., Pollack, J. M., White, C. D., Batchelor, J. H., Whelpley, C. E., ... & Adkins, C. L. (2016). Questions about questionable research practices in the field of management: A guest commentary. *Journal of Management*, *42*(1), 5-20.

Banks, G.C., Rogelberg, S.G., Woznyj, H.M. *et al.* Editorial: Evidence on Questionable Research Practices: The Good, the Bad, and the Ugly. *J Bus Psychol* 31, 323–338 (2016). https://doi.org/10.1007/s10869-016-9456-7

Bertoldo, G. (2019, November 6). Dealing with the replication crisis in psychological science: The contribution of Type M and Type S errors. DOI: https://doi.org/10.31237/osf.io/w63h7

Bertoldo, G., Zandonella Callegher, C., & Altoè, G. (2022). Designing studies and evaluating research results: Type M and Type S Errors for pearson correlation coefficient. *Meta-psychology*, 6, 1-18.

Crandall, C. S., & Sherman, J. W. (2016). On the scientific superiority of conceptual replications for scientific progress. *Journal of Experimental Social Psychology*, *66*, 93-99.

Cumming, G. (2014). The New Statistics: Why and How. *Psychological Science*, 25(1), 7–29. https://doi.org/10.1177/0956797613504966

Diedenhofen, B., & Musch, J. (2015). cocor: A comprehensive solution for the statistical comparison of correlations. *PloS one*, *10*(4), e0121945.

Fabrigar, L. R., Wegener, D. T., & Petty, R. E. (2020). A Validity-Based Framework for Understanding Replication in Psychology. *Personality and Social Psychology Review*, 24(4), 316–344. https://doi.org/10.1177/1088868320931366

Fanelli, D. (2018). Is science really facing a reproducibility crisis, and do we need it to?. *Proceedings of the National Academy of Sciences*, 115(11), 2628-2631.

Ferlazzo F. Frodi scientifiche e modelli epistemologici, in "Rassegna di Psicologia" 3/2011, pp. 5-8, doi: 10.7379/70606

Ferguson, C. J., & Brannick, M. T. (2012). Publication bias in psychological science: prevalence, methods for identifying and controlling, and implications for the use of meta-analyses. *Psychological methods*, 17(1), 120.

Fiedler, K., & Schwarz, N. (2016). Questionable research practices revisited. *Social Psychological and Personality Science*, *7*(1), 45-52.

Flake JK, Fried EI. Measurement Schmeasurement: Questionable Measurement Practices and How to Avoid Them. *Advances in Methods and Practices in Psychological Science*. 2020;3(4):456-465. doi:10.1177/2515245920952393

Funder, D. C., & Ozer, D. J. (2019). Evaluating effect size in psychological research: Sense and nonsense. *Advances in Methods and Practices in Psychological Science*, 2(2), 156-168.

Gelman, A., & Carlin, J. (2014). Beyond power calculations: Assessing type S (sign) and type M (magnitude) errors. *Perspectives on Psychological Science*, 9(6), 641-651.

Gelman, A., & Stern, H. (2006). The difference between "significant" and "not significant" is not itself statistically significant. *The American Statistician*, *60*(4), 328-331.

Gigerenzer, G. (2018). Statistical rituals: The replication delusion and how we got there. *Advances in Methods and Practices in Psychological Science*, 1(2), 198–218. DOI: https://doi.org/10.1177/2515245918771329

Hardwicke, T. E., & Wagenmakers, E. J. (2023). Reducing bias, increasing transparency and calibrating confidence with preregistration. *Nature Human Behaviour*, 7(1), 15-26.

Hensel, W.M. Double trouble? The communication dimension of the reproducibility crisis in experimental psychology and neuroscience. *Euro Jnl Phil Sci* **10**, 44 (2020). https://doi.org/10.1007/s13194-020-00317-6

Hensel, P. G. (2021). Reproducibility and replicability crisis: How management compares to psychology and economics–A systematic review of literature. *European Management Journal*, *39*(5), 577-594.

Howard, G. S., & Maxwell, S. E. (2023). ORMA: A strategy to reduce Psychology's replication problems. *New Ideas in Psychology*, *68*, 100991.

Hussey, I. (2023). A systematic review of Null Hypothesis Significance Testing, sample sizes and statistical power in research using the Implicit Relational Assessment Procedure. *Journal of Contextual Behavioral Science*.

Ioannidis, J. P. A. (2005) Why most published research findings are false. PLoS Med 2(8): e124. DOI: https://doi.org/10.1371/journal.pmed.0020124

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.

Lewis Jr, N. A. (2021). What counts as good science? How the battle for methodological legitimacy affects public psychology. *American Psychologist*, 76(8), 1323.

Maxwell, S. E., Lau, M. Y., & Howard, G. S. (2015). Is psychology suffering from a replication crisis? What does "failure to replicate" really mean?. *American Psychologist*, 70(6), 487.

Munafò, M. R., Nosek, B. A., Bishop, D. V., Button, K. S., Chambers, C. D., Percie du Sert, N., ... & Ioannidis, J. (2017). A manifesto for reproducible science. *Nature human behaviour*, *1*(1), 1-9.

Munafò, M. R., Chambers, C., Collins, A., Fortunato, L., & Macleod, M. (2022). The reproducibility debate is an opportunity, not a crisis. *BMC Research Notes*, *15*(1), 1-3.

Nieuwenhuis, S., Forstmann, B. U., & Wagenmakers, E. J. (2011). Erroneous analyses of interactions in neuroscience: a problem of significance. *Nature neuroscience*, *14*(9), 1105-1107.

Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences*, *115*(11), 2600-2606.

Nosek, B. A., Errington, T. M. (2020) What is replication? *PLoS Biol* 18(3): e3000691. DOI: https:// doi.org/10.1371/journal.pbio.3000691

Nosek, B. A., & Lakens, D. (2014). Registered reports. *Social Psychology*.

O'Hagan, A. (2019). Expert knowledge elicitation: subjective but scientific. *The American Statistician*, 73(sup1), 69-81.

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

Pashler, H., & Wagenmakers, E. J. (2012). Editors' introduction to the special section on replicability in psychological science: A crisis of confidence?. Perspectives on psychological science, 7(6), 528-530.

Pennington, *C. (2023).* A Student's Guide to Open Science: Using the Replication Crisis to Reform Psychology. *Open University Press*

Ravn, T., & Sørensen, M. P. (2021). Exploring the Gray Area: Similarities and Differences in Questionable Research Practices (QRPs) Across Main Areas of Research. *Science and Engineering Ethics*, 27(4), [40]. https://doi.org/10.1007/s11948-021-00310-z

Shrout, P. E., & Rodgers, J. L. (2018). Psychology, science, and knowledge construction: Broadening perspectives from the replication crisis. *Annual review of psychology*, 69, 487-510.

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. https://doi.org/10.1177/0956797611417632

Spitzer L, Mueller S (2023) Registered report: Survey on attitudes and experiences regarding preregistration in psychological research. *PLoS ONE* 18(3): e0281086. https://doi.org/ 10.1371/journal.pone.0281086

Tackett, J. L., Brandes, C. M., King, K. M., & Markon, K. E. (2019). Psychology's replication crisis and clinical psychological science. *Annual review of clinical psychology*, 15, 579-604.

Valentine, K.D., Buchanan, E.M., Cunningham, A., Hopke, T., Wikowsky, A. and Wilson, H. (2021), Have psychologists increased reporting of outliers in response to the reproducibility crisis?. *Soc Personal Psychol Compass*, 15: e12591. DOI: https://doi.org/10.1111/spc3.12591

Vazire, S., Schiavone, S. R., & Bottesini, J. G. (2022). Credibility beyond replicability: Improving the four validities in psychological science. *Current Directions in Psychological Science, 31*(2), 162–168. https://doi.org/10.1177/09637214211067779

Wagenmakers E-J, Dutilh G. Seven Selfish Reasons for Preregistration. 2016 Oct 31 [cited 2023 Jan 13]. In: APS Obs [Internet]. https://www.psychologicalscience.org/observer/seven%C3%A2%E2%82%AC%20selfish%C3%A2%E2%82%AC%20reasons%C3%A2%E2%82%AC%20for%C3%A2%E2% 82%AC%20preregistration

Yang, Y., Sánchez-Tójar, A., O'Dea, R. E., Noble, D. W., Koricheva, J., Jennions, M. D., ... & Nakagawa, S. (2023). Publication bias impacts on effect size, statistical power, and magnitude (Type M) and sign (Type S) errors in ecology and evolutionary biology. *BMC biology*, 21(1), 1-20.

Zondervan-Zwijnenburg, M., Van de Schoot-Hubeek, W., Lek, K., Hoijtink, H., & Van de Schoot, R. (2017). Application and evaluation of an expert judgment elicitation procedure for correlations. *Frontiers in psychology*, 8, 90.

Appendice

Appendice A: Codici per la Retrospective e Prospective Design Analysis – *retro_rho2* e *prosp_rho2*

```
# arrotondamento corretto vedi
https://stackoverflow.com/questions/12688717/round-up-from-
5
# funzione interna
round2 <- function(x, digits = 0) {  # Function to always
round 0.5 up
posneg <- sign(x)
z <- abs(x) * 10^digits
z <- z + 0.5
z <- trunc(z)
z <- z / 10^digits
z * posneg
}
####### funzione per la retrospective design analysis
retro_rho2<-function(rho1,n1,rho2,n2=NA,B=1e3,alpha=.05){
require(cocor) ; require(MASS)
if  (rho1<=rho2){
stop("First correlation must be greater than second
correlation")
}
if (is.na(n2)) n2=n1
diff_plaus=rho1-rho2
allris<-data.frame(B=1:B,rdif12=NA,p.val=NA)
for (i in 1:B){
```

```
d1<-
data.frame(mvrnorm(n=n1,mu=c(0,0),Sigma=matrix(c(1,rho1,rho
1,1),nrow=2)))
d2<-
data.frame(mvrnorm(n=n2,mu=c(0,0),Sigma=matrix(c(1,rho2,rho
2,1),nrow=2)))
d1cor<-cor(d1)[1,2]
d2cor<-cor(d2)[1,2]
d1d2cor=d1cor-d2cor
test<-cocor(~ X1 + X2 | X1 + X2,
data = list(d1, d2))
p.val=test@fisher1925$p.value
allris[i,c(2,3)]=c(d1d2cor,p.val)}
###
pow=sum(allris$p.val<alpha)/B
typeS <- length(allris$p.val[(allris$p.val<alpha) &
(allris$rdif12<0)]) / sum(allris$p.val<alpha)
if (is.na(typeS)) typeS<-0
typeM <- mean(abs(allris$rdif12[allris$p.val<alpha]))/
diff_plaus
return(list(power=pow,typeS=typeS,typeM=typeM))
}
####### funzione per la prospective design analysis
prosp_rho2<-
function(rho1,rho2,alpha=.05,power=.80,B=1e3,rangen=c(10,20
00),tol = .01,iter=20,verbose=FALSE)
{
if  (rho1<=rho2){
stop("First correlation must be greater than second
correlation")
}
j=0
diff_plaus=rho1-rho2
```

```
# check for max N
n1<-n2<-rangen[2]
vals<-
retro_rho2(rho1=rho1,rho2=rho2,n1=n1,n2=n2,alpha=alpha,B=B)
if (vals$power<power)
{ cat(paste0("Actual power = ", vals$power, " with n = ",
rangen[2], " (per group), " ),"\n")
cat(paste0("   try to set minimum of rangen to ",
rangen[2], " and increase maximum of rangen.","\n"))
return(vals)
}
# propspective
n_seq <- seq( rangen[1], rangen[2], by = 1 )
n_target <- round2(median(n_seq))
find_power=FALSE
#
while( (!find_power) ) {
if (verbose!=FALSE) {cat("Estimating power with n per group
=",n_target,"\n")}
est_P<-
retro_rho2(rho1=rho1,rho2=rho2,n1=n_target,n2=n_target,alph
a=alpha,B=B)
est_power <- est_P$power
if (verbose!=FALSE) {cat("Estimated power est_power
=",est_power,"\n")}
if ( (est_power<=(power+tol)) & (est_power>(power-tol)) ) {
find_power <- TRUE
} else {
if (length(n_seq)==1) {
print(n_seq)
stop(" ")
}
```

```
if ( est_power > (power-tol) ) {
(n_seq <- seq( min(n_seq), n_target, by = 1))
(n_target <- round2(median(n_seq)))
} else {
(n_seq <- seq( n_target, max(n_seq), by = 1))
(n_target <- round2(median(n_seq)))
}
}
j=j+1
if (j==iter)
{ cat(paste0("The algorithm did not achieve convergence
with iter =  ", j,"\n"))
cat(paste0("Try to increase tol and/or iter" ,"\n"))
return(NA)}
}
ris <- list(power=power,est_power= est_power, n_group=
n_target, typeM=est_P$typeM, typeS=est_P$typeS)
return(ris)
}
```