**UNIVERSITÀ DEGLI STUDI DI PADOVA**

**Dipartimento di Dipartimento di Psicologia dello Sviluppo e della Socializzazione**

**Corso di laurea in Scienze e Tecniche Psicologiche**

# Test statistici di equivalenza in psicologia

**Statistical equivalence tests in psychology**

***Relatore***
**Prof. Livio Finos**

***Laureanda: Ilaria Santini***
***Matricola: 2021964***

Anno Accademico 2022/2023

# Indice

# 1. Contestualizzazione sulle problematiche di ricerca in psicologia

## 1.1 Crisi della replicabilità

Uno degli aspetti più importanti per la validità dei risultati in ricerca scientifica riguarda la loro replicabilità. Si tratta della ripetizione di una specifica procedura sperimentale applicata su un nuovo campione diverso da quello dell'esperimento originale con l'obiettivo di valutare se vengono ottenuti gli stessi risultati per verificarne l'affidabilità (Nosek et al., 2022). Più volte viene ottenuto lo stesso risultato, più questo ottiene evidenze empiriche a suo favore. Si tratta dunque di un caposaldo fondamentale del metodo scientifico senza il quale non si possono trarre conclusioni affidabili. Si distingue dalla riproduzione perché questa viene effettuata sui medesimi dati che sono stati raccolti nell'esperimento originale.

Nell'ambito della ricerca psicologica, così come in altri campi, da anni i ricercatori hanno a che fare con la problematica della crisi della replicabilità dei risultati che riguarda la difficoltà, o anche impossibilità, di replicare alcuni studi, causando un generale dibattito sulla effettiva credibilità dei loro metodi e dei loro risultati. Uno dei motivi principali sembra essere il grande utilizzo in campo psicologico di Pratiche di Ricerca Discutibili (PDR). Si tratta di procedure che si trovano a cavallo tra pratiche accettabili e non accettabili (O'Boyle & Götz, 2022). Rientrano in questa categoria tutti quei comportamenti che sono volti all'aumentare la probabilità del ricercatore di confermare la sua ipotesi di partenza, così come il raccogliere dati finchè non si ottengono risultati significativi e il riportare solo le ipotesi che sono state confermate (O'Boyle & Götz, 2022). Dagli studi condotti (John et al., 2012) emerge un'alta percentuale di psicologi ricercatori che hanno ammesso di aver utilizzato PDR nel corso delle loro ricerche. Oltre ai casi di falsificazione dei dati, che comunque restano casi isolati, una pratica che pare essere molto comune in ambito psicologico è l'Harking (Hypothesizing After the Results are Known) che consiste nel trasformare una ricerca esplorativa, quindi atta ad analizzare dei dati, in una confermativa, che prevede un'ipotesi di partenza da verificare, nel corso dello studio (O'Boyle & Götz, 2022).

Emerge, da una metanalisi condotta nel 2012 (Makel et al., 2012) che solo l'1,07% degli studi pubblicati dal 1900 al 2012 sono replicazioni di esperimenti condotti in precedenza. Anche se questo dato sembra essere in aumento dall'inizio del XXI secolo, trattandosi di una cifra molto

bassa, è chiaro come la grande maggioranza di risultati, essendo non replicati, sono isolati e non possono essere considerati significativi.
Sembra essere molto comune, inoltre, il fatto di accettare scoperte come tali basandosi solo su un singolo studio che ha affermato tale conclusione esclusivamente sul valore p (Ioannidis, 2005). In aggiunta a ciò, si unisce una problematica che sembra essere molto comune nel campo della ricerca, ossia il mal interpretare il concetto di *p-value* nei risultati dei test ad ipotesi nulla (Colquhoun, 2014), di cui si parlerà nei prossimi capitoli.

Dal momento che non è possibile ottenere una replicazione esatta, uno studio si definisce replicato quando la differenza dei risultati dello studio originale e della sua replicazione viene considerata uguale a zero (Nosek et al., 2022). Inoltre, essendo chiaro che la cosiddetta "replicazione esatta", ossia la possibilità di poter replicare le esatte condizioni di un contesto specifico, non sia possibile, la replicabilità di uno studio si basa sul numero di prove effettuate a suo favore (Nosek et al., 2022).
Per replicare uno studio quando si sospettano falsi positivi sono stati introdotti i test di equivalenza. Essi risultano essere una possibile soluzione alla crisi di replicazione perché si tratta di un approccio statistico per valutare se un effetto è sufficientemente piccolo da concludere che lo studio di partenza non sia replicabile (Lakens, 2022). Inoltre, è possibile utilizzare per testare se lo studio originale e la sua replicazione possono essere considerati equivalenti entro un certo intervallo. Nel Capitolo 4 verrà presentata un'applicazione per quest'ultimo caso.

## 1.2 Test delle ipotesi: NHST

Questo sotto capitolo sarà dedicato ad un'ulteriore problematica legata all'attuale contesto di ricerca in psicologia: l'NHST (*Null Hypothesis Significance Testing*). In particolare, verrà evidenziato come si è sviluppato, la sua applicazione e alcuni dei suoi principali punti più criticati.

L'NHST è la procedura più comune per testare i dati e si presenta come un "modello ibrido" tra la teoria di Fisher e quella di Neyman-Pearson (Perezgonzalez, 2015). Analizzando singolarmente i due approcci, quello di Fisher e quello di Neyman-Pearson, notiamo come in realtà essi siano concettualmente discordanti su vari punti e questa incompatibilità dà luogo ad alcuni dei fraintendimenti e malintesi che vengono riscontrati in un gran numero di ricerche e

studi, aumentando il numero di risultati non replicabili nella letteratura (Perezgonzalez, 2015), di cui si ne è parlato nel Capitolo 1.1. Nonostante le due teorie usino gli stessi strumenti e portino agli stessi risultati statistici, la differenza principale si riscontra nel modo in cui vengono interpretati i risultati e come viene impostata filosoficamente la ricerca (Perezgonzalez, 2015).
Innanzitutto, per Fisher non è specificata una precisa ipotesi alternativa, mentre in Neyman-Pearson è necessaria e va definita a priori, da cui vengono poi calcolati la potenza del test, l'errore del I tipo (falso positivo, α) e l'errore del II tipo (falso negativo, β). Questi ultimi tre concetti non vengono presi in considerazione nella teoria di Fisher e vengono infatti introdotti successivamente dai due studiosi. Per Fisher la significatività statistica è rappresentata dal valore *p* che corrisponde alla probabilità di ottenere dati corrispondenti al campione osservato o più estremi, se l'ipotesi nulla è vera. Di conseguenza, minore sarà il *p-value* maggiore sarà l'evidenza statistica contro $H_0$.
Nell'approccio di Neyman-Pearson, invece, la significatività del test equivale a α, ossia l'errore del I tipo, definito come la probabilità di rifiutare $H_0$ quando questa è vera. Un ulteriore concetto introdotto da Neyman e Pearson è l'errore del II tipo (β), ossia la probabilità di non rifiutare $H_0$ quando questa è falsa. Ne consegue che la potenza di un test equivalga alla capacità di riconoscere che $H_0$ sia falsa quando questa è effettivamente falsa e viene espressa come $1 - \beta$. Secondo questo approccio si può rifiutare o meno l'ipotesi nulla accettando o meno quella alternativa ($H_1$), mentre per Fisher il rifiuto di $H_0$ è soggettivo e riguarda il livello di significatività dato al *p-value* che non viene specificato a priori, così come l'eventuale ipotesi alternativa (Perezgonzalez, 2015).
Risultano ora più chiari alcuni aspetti metodologici legati alla "nascita" dell'NHST che evidenziano due teorie alla base in netto contrasto tra loro (Perezgonzalez, 2015).

Normalmente, quando siamo interessati a verificare l'esistenza di un effetto, usiamo il cosiddetto test di significatività dell'ipotesi nulla (Walker & Nowacki, 2011), il *Null Hypothesis Significance Testing*, nel quale impostiamo appunto due ipotesi opposte: l'ipotesi alternativa ($H_1$) e l'ipotesi nulla ($H_0$). La prima riguarda ciò che vogliamo dimostrare, ossia la presenza dell'effetto, ed equivale quindi alla nostra ipotesi di ricerca. L'ipotesi nulla, invece, è il suo esatto opposto, e quindi ciò che vogliamo confutare. L'obiettivo di questi studi comparativi è quello di raccogliere abbastanza evidenze per poter concludere l'esistenza di una differenza significativamente diversa da zero.
Da due campioni ($n_1$ e $n_2$), estratti da popolazioni con ugual distribuzione, vengono raccolte rispettivamente osservazioni indipendenti tra loro. La procedura più comune per valutare la

presenza di una differenza tra le due medie delle popolazioni ($\mu_1$ e $\mu_2$) è, appunto, il test di significatività dell'ipotesi nulla (NHST). Questo approccio consiste nello specificare due ipotesi (***Figura 1***):

a. L'ipotesi nulla ($H_0$) che riguarda l'assenza di un effetto, ossia che la differenza tra le due medie sia uguale a zero.

$$H_0: \mu_1 - \mu_2 = 0$$

anche espressa come

$$H_0: \mu_1 = \mu_2$$

b. L'ipotesi alternativa ($H_1$), che corrisponde alla nostra ipotesi di ricerca, che riguarda qualunque effetto diverso da zero.

$$H_1: \mu_1 - \mu_2 \neq 0$$

anche espressa come

$$H_1: \mu_1 \neq \mu_2$$

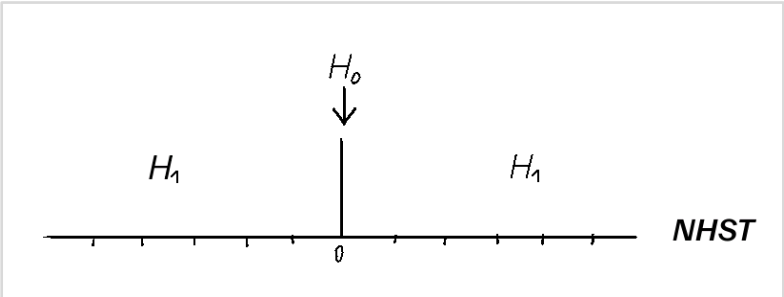


***Figura 1***. *Visualizzazione dell'ipotesi nulla ($H_0$) e dell'ipotesi alternativa ($H_1$) nel NHST.*

Successivamente, viene calcolata la statistica t-test sulla base dei dati campionati per testare le due medie. Nei test di significatività dell'ipotesi nulla si testa se esiste un effetto o se i dati ottenuti sono esclusivamente dovuti al caso (Lakens, 2022). Quest'ultima possibilità viene espressa tramite un valore *p* (*p-value*) calcolato che corrisponde dunque alla probabilità di osservare, assumendo che l'ipotesi nulla sia vera, dati del campione o valori più estremi (***Figura 2***).

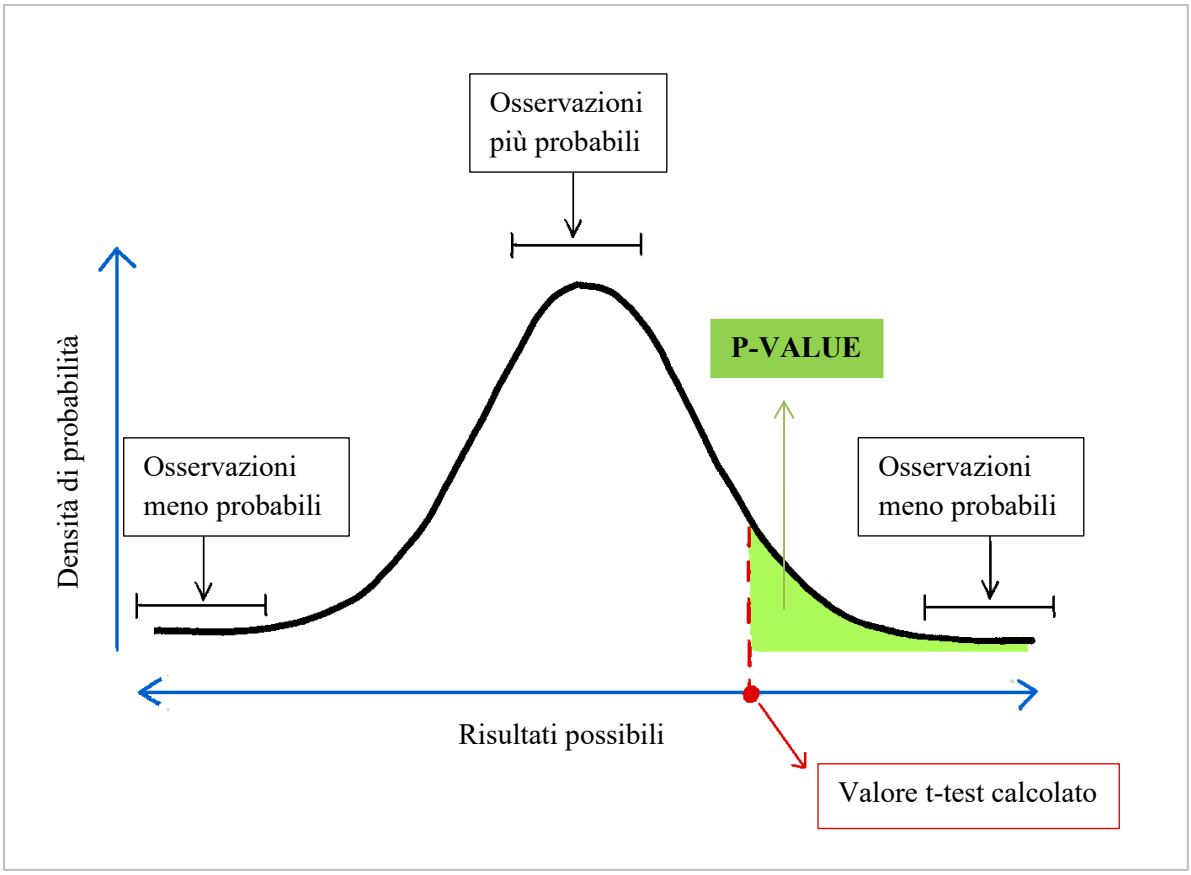


***Figura 2.*** *Visualizzazione del p-value (zona evidenziata in verde) nella distribuzione di probabilità dei possibili valori.*

Una volta ottenuto il valore t-test verrà calcolato, dunque, il p-value. Questa probabilità viene confrontata con il livello di significatività (α) che è la probabilità di rifiuto dell'ipotesi nulla quando è vera, viene impostata a priori e corrisponde solitamente al 5% (0.05).

Se il p-value risulta essere inferiore a α (grafico in alto, ***Figura 3***), è possibile rifiutare $H_0$ a favore di $H_1$. Se, invece, risulta essere maggiore o uguale a α (grafico in basso, ***Figura 3***), non è possibile rifiutare l'ipotesi nulla perché non ci sono prove sufficienti a favore dell'ipotesi alternativa. In altre parole, un risultato può essere definito significativo se il valore p è inferiore al livello di significatività (α).

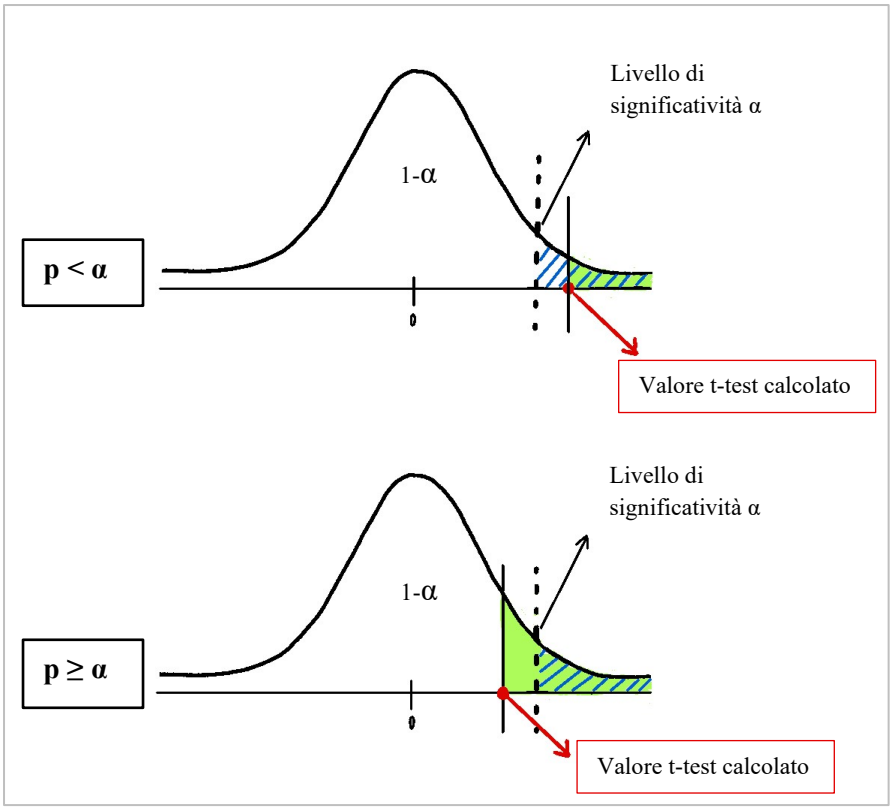


***Figura 3**. Rappresentazione aree p-value (in verde) e α (linee blu) nei due casi possibili (p-value < α o p-value > α).*

Innanzitutto, è importante sottolineare che l'ipotesi alternativa non viene realmente specificata, perché corrisponde "a tutto ciò che non è zero". Di conseguenza, quando si effettua un NHST, non viene delineato che valore ci si aspetta di ottenere dai dati raccolti, ossia non viene quantificato l'effetto oggetto di studio. Ne consegue che l'ipotesi alternativa può essere falsificata solo ottenendo un effetto uguale a zero (Lakens, 2019) e ciò non risulta possibile perchè, data la variabilità intrinseca dei dati e delle misurazioni, un valore non potrà mai risultare esattamente zero. Infatti, anche se raccogliessimo infiniti dati di un fenomeno in cui non esiste un effetto reale ("vero"), i dati varierebbero attorno allo zero a causa di variabili casuali ed errori di misurazione. Questa precisazione evidenzia, dunque, perché l'NHST sia un test non falsificabile e ciò lo rende meno "valido" a livello scientifico[1]. Sarebbe più opportuno prendere in considerazione, invece che "l'esatto valore zero", un intervallo attorno allo zero.

Riassumendo, il test di significatività dell'ipotesi nulla si basa su un concetto che viene spesso mal interpretato, ossia il p-value e i ricercatori sembrano eccessivamente focalizzati sull'ottenere risultati statisticamente significativi invece che su un'effettiva analisi approfondita dei dati. Viene data troppa enfasi sulla significatività (statistica) e meno su quella pratica (rilevanza). La

[1] Per un breve approfondimento vedere la voce "principio della falsificabilità" nel *dizionario di filosofia* del 2009 della Treccani.

prima corrisponde all'osservare un effetto statisticamente diverso da zero e che sia poco probabile che sia dovuto al caso, mentre la seconda riguarda la rilevanza da un punto di vista pratico. Un risultato statisticamente significativo potrebbe comunque essere troppo piccolo per avere una rilevanza clinica.

Inoltre, l'NHST non specifica l'ipotesi alternativa in modo dettagliato e, essendo l'ipotesi nulla corrispondete al valore zero, non risulta essere un test falsificabile. Come ne parleremo nel prossimo sotto capitolo, il p-value di questa tipologia di test non permette di accettare l'ipotesi nulla, cioè non consente di testare l'assenza di un effetto ma solo di rifiutare o non rifiutare $H_0$.

## 1.3 Problematiche legate al testare l'assenza di un effetto

Come affermato in precedenza, se lo scopo di uno studio è quello di valutare la presenza di un effetto, tramite il test di significatività dell'ipotesi nulla, i ricercatori possono rifiutare $H_0$ quando il *p-value* risulta essere inferiore ad α. Nel momento in cui, però, esso risulta essere maggiore, non possiamo né rifiutarla né accettarla. Molti ricercatori commettono l'errore logico di concludere che non ci sia effetto se non si può rifiutare l'ipotesi nulla (Lakens, 2017). In altre parole, non è possibile utilizzare un NHST per affermare un'equivalenza ("uguale a zero") (Walker & Nowacki, 2011).

Innanzitutto, il primo problema che si pone riguarda il fatto che l'ipotesi di ricerca ($H_1$), quella che il test "valuta", non corrisponde all'assenza dell'effetto, ma bensì al suo esatto opposto. In sostanza, come specificano Walker e Nowacki (2011), un risultato significativo in un classico test comparativo permette di concludere la presenza di un effetto, ma un risultato non significativo (per esempio un *p-value* $> 0.05$ per un $\alpha = 0.05$) permette esclusivamente di arrivare alla conclusione che l'equivalenza non può essere esclusa. Questo non implica in alcun modo che possa essere accettata. Inoltre, essendo l'ipotesi nulla di un NSHT un effetto uguale a zero, non viene incluso il concetto di margine di equivalenza, perchè non definito (Walker & Nowacki, 2011).

Se non dimostro la presenza di un effetto questo non vuol dire che ho dimostrato la sua assenza (Lakens, 2022). Di conseguenza, utilizzare la non significatività di un test ad ipotesi nulla per affermare che non ci sia un effetto non è né statisticamente né logicamente corretto (Lakens et al., 2020). Essa, infatti, ci dice esclusivamente che non possiamo rifiutare $H_0$ perché il valore *p* di un NHST può darci solo questa risposta (Lakens, 2022).

Non è possibile, dunque, concludere l'assenza di un effetto se non riesco a dimostrarne la sua presenza (Lakens, 2022).

L’insieme di criticità legate alla replicabilità dei risultati e all’utilizzo degli NHST, così come alle problematiche legate al testare l’assenza di un effetto, è una delle principali cause che porta a risultati non validi in ambito psicologico. Una soluzione a queste criticità risultano essere i test di equivalenza.

## 2. Test di equivalenza

Ci sono varie circostanze in cui un ricercatore potrebbe essere interessato a verificare che non ci sia una differenza significativa tra due fenomeni, per esempio se vogliamo sapere se un tipo di intervento ha lo stesso esito di un altro oppure nel caso in cui vogliamo replicare uno studio in cui abbiamo motivo di credere che l'effetto ottenuto sia un falso positivo (errore di I tipo) (Lakens, 2022). Se siamo quindi intenzionati a testare che un effetto sia zero, la domanda di ricerca e il procedimento cambiano. In particolare, rispetto ad un normale test delle ipotesi, l'ipotesi nulla e l'ipotesi alternativa si invertiranno (***Tabella 1***).

***Tabella 1***. *Ipotesi associate in base alla tipologia di ricerca quando si compara una nuova terapia ad una già esistente (originale) in base all'efficacia (Walker & Nowacki, 2011).*

| ***Tipo di ricerca*** | ***Ipotesi nulla ($H_0$)*** | ***Ipotesi di ricerca ($H_1$)*** |
|---|---|---|
| ***Comparativa tradizionale*** | *Non c'è differenza tra le due terapie* | *C'è differenza tra le due terapie* |
| ***Equivalenza*** | *Le terapie non sono equivalenti* | *La nuova terapia è equivalente all'originale* |

Presupponiamo di voler testare se una nuova terapia più economica abbia la stessa efficacia di quella che viene attualmente utilizzata in quel campo. Invece di avere come ipotesi di ricerca una differenza tra medie, avremo una differenza uguale a zero. Come abbiamo già specificato nel capitolo 1.2, non è possibile dimostrare che un effetto sia esattamente uguale a zero, allora testeremo che la differenza rientri all'interno di un margine che definiremo "equivalente a zero" (Walker & Nowacki, 2011).

I test di equivalenza sono stati introdotti inizialmente nelle scienze farmaceutiche (Lakens, 2022) e vengono utilizzati per testare se un nuovo farmaco possa essere considerato equivalente ad uno già esistente, caratteristica nota con il termine di "bioequivalenza".
Sarebbe opportuno che venissero maggiormente utilizzati in ambito di ricerca psicologica per vari motivi che Lakens espone (2017). In primo luogo, l'utilizzo dei test di equivalenza porterebbe ad una maggiore conoscenza del valore *p*, diminuendo la probabilità che questo venga interpretato in modo scorretto e, di conseguenza, si eviterebbe che i ricercatori traggano la

conclusione dell'assenza di un effetto basandosi esclusivamente sulla non significatività dei risultati ($p\text{-}value > \alpha$). Un altro beneficio riguarda più in generale la possibilità di testare l'assenza di effetti significativi dal momento in cui i classici test di significatività dell'ipotesi nulla non lo permettono. Infine, i test di equivalenza impongono ai ricercatori di indagare maggiormente sull'effetto che si aspettano di osservare e quindi anche sulla differenza minima che è ritenuta significativa.

## 2.1 Definizione e margini di equivalenza

Nei test di equivalenza, invece che rifiutare $H_0$ quando è diverso da zero, la rifiutiamo quando l'effetto sta all'interno di un intervallo attorno allo zero che consideriamo equivalente (***Figura 4***). In sostanza, si valuta se l'effetto è abbastanza estremo da essere considerato significativo, ossia non nullo (Lakens, 2022) e, quindi, invece di voler confutare l'ipotesi che un effetto sia uguale a zero, come facciamo nei NHST, vogliamo confutare l'ipotesi che l'effetto sia un range di valori che si discostano da zero.

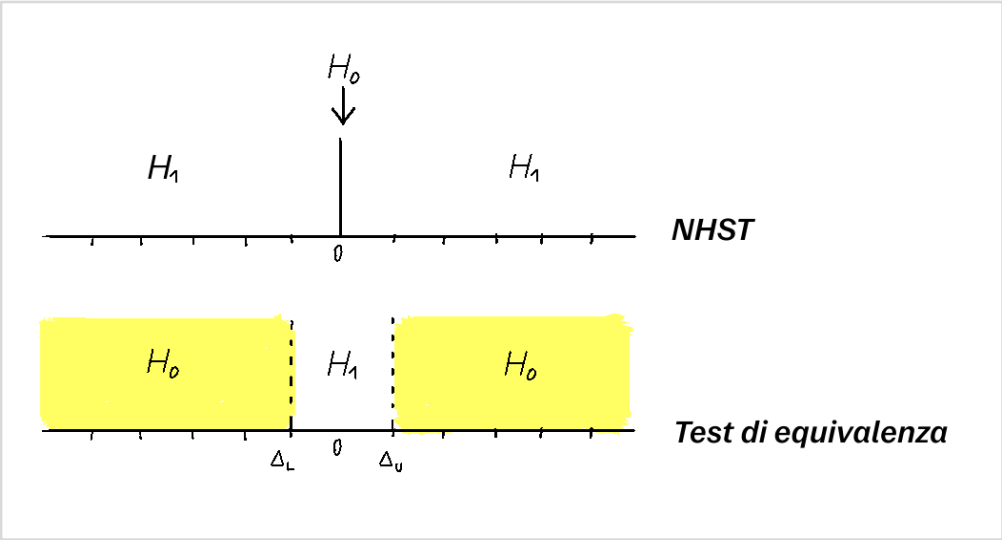


***Figura 4.*** *Visualizzazione delle ipotesi nulle ($H_0$) e delle ipotesi alternative ($H_1$) rispettivamente per il test della significatività dell'ipotesi nulla (NHST) e il test di equivalenza.*

Il primo passo da compiere è quello di definire quali sono questi valori, cioè di quantificare l'intervallo entro il quale l'effetto non viene considerato significativo.
Innanzitutto, dal momento che un effetto non sarà mai uguale a zero, bisogna specificare come dovrebbe apparire per essere considerato sufficientemente piccolo.

È necessario quindi stabilire un range di equivalenza, ossia un intervallo entro il quale l'effetto viene considerato "praticamente equivalente" a zero.

Si presuppone che lo scopo di ogni ricerca sia sempre quello di arricchire la conoscenza e quindi è necessario che il valore ottenuto, per essere definito significativo, sia giustificato in base al contesto di riferimento specificando perché quell'effetto avrà una rilevanza pratica.
Nei test di significatività delle ipotesi definiamo $H_0$ come nulla e $H_1$ come tutto il resto ma nei test di equivalenza essendo $H_0$ un range di valori e $H_1$ un intervallo attorno allo zero, invertiamo l'ipotesi nulla con l'ipotesi di ricerca abituali. Emerge, quindi, la necessità di definire l'ipotesi alternativa in modo più dettagliato (Lakens et al., 2020) e cioè di giustificare la "soglia" oltre la quale l'effetto viene ritenuto significativo.
Lo Small Effect Size Of Interest (SESOI) è l'effetto minimo teoricamente interessante che interessa ai ricercatori e che ha una rilevanza pratica (Anvari & Lakens, 2021). Il SESOI può essere stabilito in base a criteri oggettivi come, per esempio, in base a teorie che determinano una "dimensione soglia" che sia clinicamente rilevante. In caso contrario, si tratterà di SESOI stabiliti in base a criteri soggettivi e, come tali, potrebbero non essere condivisi da tutti i ricercatori. È fondamentale, quindi, giustificare l'effetto minimo che si ritiene significativo, in modo tale da dar la possibilità di interpretare lo studio e di comprendere l'interpretazione data a tali risultati. L'importanza e le possibili giustificazioni del SESOI verranno approfondite nel Capitolo 3 perché emerge prima la necessità di approfondire l'applicazione dei test di equivalenza di cui se ne parla nel prossimo sotto capitolo (2.2).

## 2.2 Procedura TOST

L'approccio più comune per i test di equivalenza è il TOST (*Two One-Sided Test*) e consiste nell'esaminare se la stima (l'intervallo di confidenza) dell'effetto osservato (per esempio impostato al 90%) cade tra $\Delta_L$ e $\Delta_U$ (ossia i limiti di equivalenza) e concludere l'equivalenza se questo intervallo non contiene nessuno dei limiti di equivalenza (Lakens, 2017).
La procedura TOST implica, quindi, di applicare due test unilaterali per verificare se i dati osservati sono maggiori di un intervallo superiore o minori di un intervallo inferiore (Lakens, 2017):

$$t_L = \frac{\overline{M_1} - \overline{M_2} - \Delta_L}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \qquad t_U = \frac{\overline{M_1} - \overline{M_2} - \Delta_U}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$$

$$\sigma = \sqrt{\frac{(n_1 - 1)sd_1^2 + (n_2 - 1)sd_2^2}{n_1 + n_2 - 2}}$$

(dove *M* corrisponde alla media di ogni campione, $\sigma$ è la deviazione standard, *n* è la numerosità del campione e *t* è il test statistico unilaterale).

In primo luogo, dunque, è necessario specificare $\Delta_L$ e $\Delta_U$ sulla base dello Small Effect Size Of Interest che abbiamo stabilito a priori e, di conseguenza, le ipotesi nulle saranno due: la prima ($H_{01}$) che l'intervallo sia inferiore o uguale a $\Delta_L$ e la seconda ($H_{02}$) che sia maggiore o uguale a $\Delta_U$ (Lakens et al., 2018) (***Figura 5***).

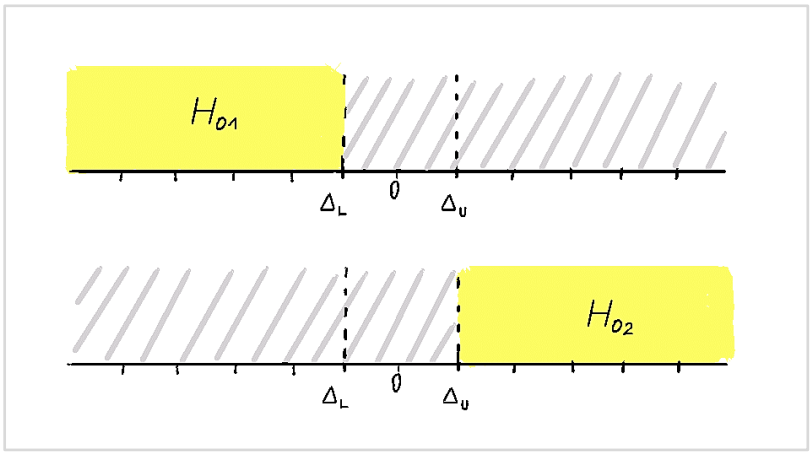


***Figura 5.***
*Visualizzazione delle ipotesi nulle dei due test unilaterali.*

Ne consegue che l'ipotesi alternativa sia che la stima dell'effetto cada tra i limiti di equivalenza. Se entrambi i test unilaterali sono significativi ($\Delta_L < \Delta < \Delta_U$), ossia se i valori *p* di entrambi i test unilaterali sono significativi, seguendo l'approccio di Neyman-Pearson, possiamo rifiutare entrambe le ipotesi nulle (Lakens, 2017) e concludere che l'effetto è troppo piccolo per essere considerato significativo nel nostro intervallo (Lakens, 2022).

Minori saranno i limiti di equivalenza, maggiore sarà la numerosità campionaria necessaria per ottenere risultati significativi (Lakens, 2017). In altre parole, impostando un SESOI molto basso, l'intervallo risulterà essere più vicino a zero e, di conseguenza, aumenterà il numero di dati necessari per concludere l'equivalenza.

È importante sottolineare che rifiutando $H_0$ non si può comunque concludere che non ci sia effetto, ma solamente che, per i nostri limiti di equivalenza, l'effetto viene considerato troppo piccolo per essere significativo. La dicitura corretta risulta essere quindi che i valori risultano "statisticamente equivalenti" per quell'intervallo (Lakens et al., 2018).

### 2.2.1 Possibili esiti e interpretazioni dei risultati

È possibile applicare sia il test della significatività dell'ipotesi nulla sia i test di equivalenza, in primo luogo per concludere l'assenza di un effetto se l'NHST risulta non significativo e quindi poter dimostrare l'equivalenza e in secondo luogo per evitare che si concluda erroneamente l'assenza di un effetto quando il valore *p* dell'NHST risulta superiore ad alfa (Lakens, 2017). In altre parole, combinando i due test è possibile evitare che risultati statisticamente significativi vengano direttamente considerati, senza un'opportuna analisi, anche come praticamente significativi (Lakens, 2022), perché "diverso da zero" non equivale per forza a "sufficientemente grande da avere una rilevanza pratica".

Nei test di equivalenza, infatti, non solo si possono rilevare effetti staticamente inferiori al SESOI ma anche effetti statisticamente diversi da zero. Questo permette di trarre conclusioni più affidabili sia sull'esistenza dell'effetto, sia se questo è abbastanza grande da essere considerato significativo (Lakens, 2022). Un NHST test può concludere che vi sia un effetto, perchè diverso da zero, ma con l'utilizzo di un test di equivalenza è possibile valutare se quell'effetto è abbastanza grande da essere considerato significativo. Allo stesso modo si può verificare se un effetto ritenuto non significativo sia effettivamente considerabile equivalente a zero.

Lakens (2022) presenta un esempio esaustivo per l'utilizzo dei test di equivalenza. Un gruppo di ricercatori chiesero ai partecipanti di trasportare scatole pesanti con il fine di variare il loro livello di fatica. L'ipotesi di partenza era che la fatica non alterasse le emozioni dei partecipanti. Non vengono rilevate differenze significative nel tono dell'umore tra il gruppo sperimentale ($m_1$) e il gruppo di controllo ($m_2$):

$m_1 = 4.55, sd_1 = 1.05, n_1 = 15$;

$m_2 = 4.87, sd_2 = 1.11, n_2 = 15$;

Secondo il test di significatività dell'ipotesi nulla non vengono rilevate differenze tra le due condizioni:

$t = -0.81, p = .42$ con una differenza tra le medie di $-0.32$

e in ricercatori concludono che la differenza non è abbastanza significativa.

A questo risultato dell'NHST applichiamo dunque il test di equivalenza per verificare se -0.32 è effettivamente abbastanza piccolo da essere considerato equivalente a zero per quell'intervallo, utilizzando la procedura TOST.

È necessario innanzitutto specificare quale sia il valore oltre il quale l'effetto sarebbe stato ritenuto significativo (Lakens, 2022). Supponiamo che, in questo esempio, il "valore soglia" e i limiti di equivalenza siano -0.5 e 0.5.

In ambito applicativo, è possibile eseguire la procedura TOST con il pacchetto "TOSTER".

In questo modo è possibile visualizzare le due distribuzioni (i due t-test unilaterali) dei limiti dell'intervallo di equivalenza (impostato a -0.5 e 0.5) e le aree di rifiuto per ogni distribuzione (aree verdi, **Figura 6**).

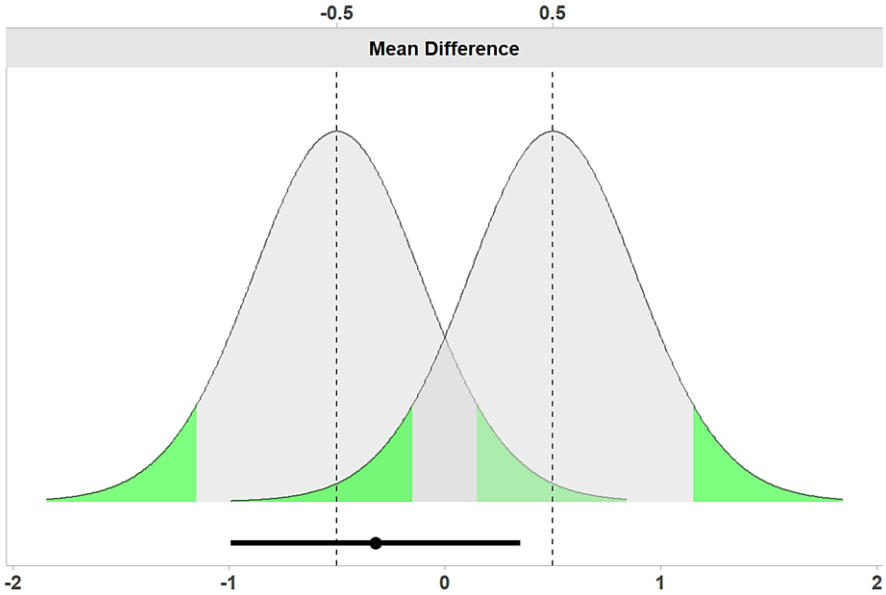


***Figura 6**. Differenza tra le medie dei due campioni e il suo rispettivo intervallo di confidenza posto al di sotto delle due distribuzioni dei limiti di equivalenza (0.5 e -0.5).*

*(Lakens,2022)*

Al di sotto delle due distribuzioni è rappresentato l'intervallo di confidenza attorno al valore -0.32.

Applicando il pacchetto TOSTER otterremo:

```
TOSTER::tsum_TOST(m1 = 4.55,
                  m2 = 4.87,
                  sd1 = 1.05,
                  sd2 = 1.11,
                  n1 = 15,
                  n2 = 15,
                  low_eqbound = -0.5,
                  high_eqbound = 0.5)
```

*(Lakens, 2022)*

```
Welch Modified Two-Sample t-Test

The equivalence test was non-significant, t(27.91) = 0.456, p = 3.26e-01
The null hypothesis test was non-significant, t(27.91) = -0.811, p = 4.24e-01
NHST: don't reject null significance hypothesis that the effect is equal to zero
TOST: don't reject null equivalence hypothesis

TOST Results
                 t    df p.value
t-test     -0.8111 27.91   0.424
TOST Lower  0.4563 27.91   0.326
TOST Upper -2.0785 27.91   0.023

Effect Sizes
               Estimate     SE              C.I. Conf. Level
Raw             -0.3200 0.3945 [-0.9912, 0.3512]         0.9
Hedges's g(av)  -0.2881 0.3930 [-0.8733, 0.3021]         0.9
Note: SMD confidence intervals are an approximation. See vignette("SMD_calcs").
```

*(Lakens, 2022)*

Le diciture "TOST Lower" e "TOST Upper" corrispondono ai risultati dei due test unilaterali. In questo caso, possiamo rifiutare l'esistenza di effetti superiori a 0.5 ($\Delta_U$) ma non possiamo rifiutare effetti inferiori a -0.5 ($\Delta_L$). È necessario che entrambi i test unilaterali risultino significativi affinché il test di equivalenza sia significativo e che quindi entrambe le ipotesi $H_{01}$ e $H_{02}$ devono essere rifiutate (***Figura 7***).

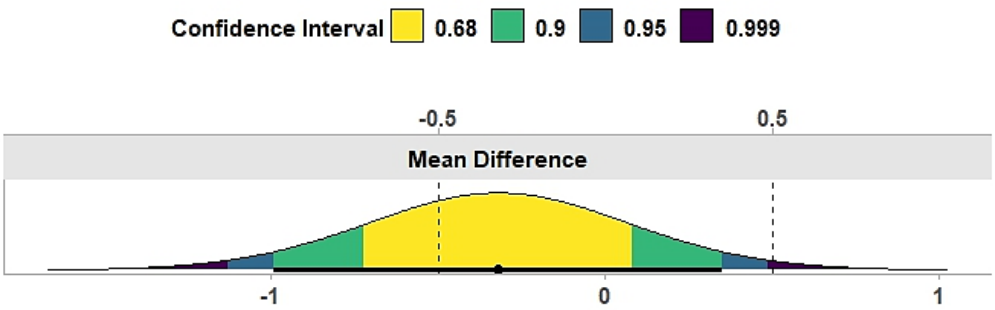


***Figura 7****. Visualizzazione degli intervalli di confidenza attorno alla differenza delle medie.*

Sia il test di equivalenza che il NHST sono risultati non significativi e quindi non possiamo concludere né che ci sia equivalenza né che l'effetto è abbastanza grande da essere significativo. Questo è un esempio di risultato inconclusivo perché, come sottolinea Lakens (2022), non è possibile sapere se effettivamente l'effetto è troppo piccolo da essere significativo o se lo studio non è stato in grado di rilevare un effetto realmente esistente (errore del II tipo).

Se lo stesso studio fosse fatto su un campione più grande, per esempio 200 partecipanti per gruppo, i risultati cambierebbero.

```
Welch Modified Two-Sample t-Test

The equivalence test was significant, t(396.78) = 1.666, p = 4.82e-02
The null hypothesis test was significant, t(396.78) = -2.962, p = 3.24e-03
NHST: reject null significance hypothesis that the effect is equal to zero
TOST: reject null equivalence hypothesis

TOST Results
                t    df p.value
t-test     -2.962 396.8   0.003
TOST Lower  1.666 396.8   0.048
TOST Upper -7.590 396.8 < 0.001

Effect Sizes
               Estimate    SE               C.I. Conf. Level
Raw             -0.3200 0.108 [-0.4981, -0.1419]         0.9
Hedges's g(av)  -0.2956 0.104 [-0.4605, -0.1304]         0.9
Note: SMD confidence intervals are an approximation. See vignette("SMD_calcs").
```

*(Lakens, 2022)*

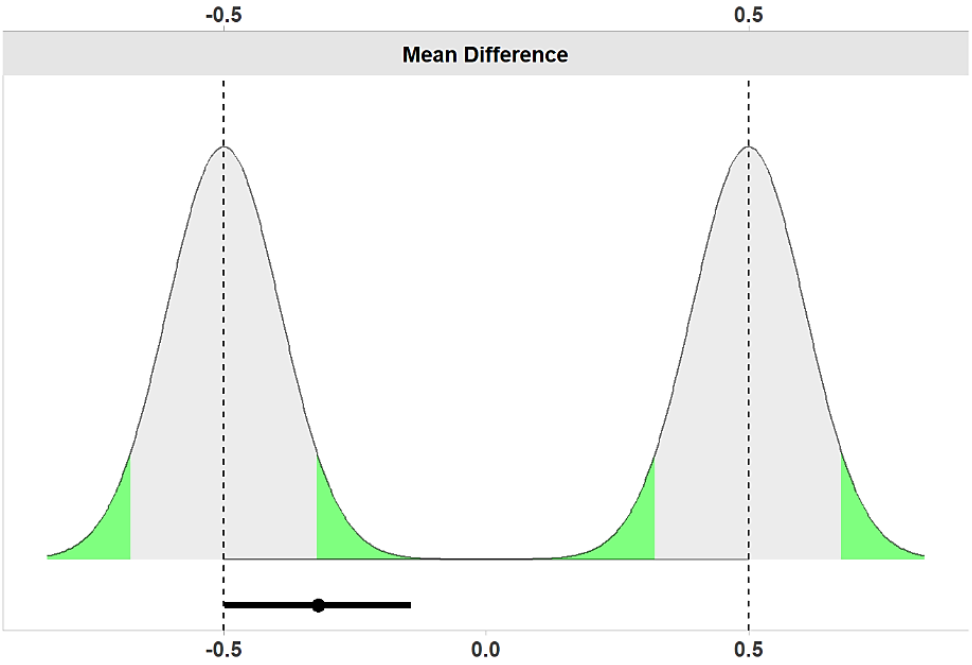


***Figura 8***.
*Differenza tra le medie dei due campioni e il suo rispettivo intervallo di confidenza posto al di sotto delle due distribuzioni dei limiti di equivalenza (0.5 e -0.5).*

*(Lakens, 2022)*

Per evitare di ottenere risultati inconclusivi risulta quindi necessario raccogliere un campione sufficientemente grande (Lakens, 2022).

In questo caso entrambi i test risultano significativi:

a. Il NHST risulta significativo perché l'intervallo di confidenza (impostato al 90%) esclude lo zero (***Figura 9***)
b. Il test di equivalenza risulta significativo perché l'intervallo di confidenza cade nei limiti di equivalenza

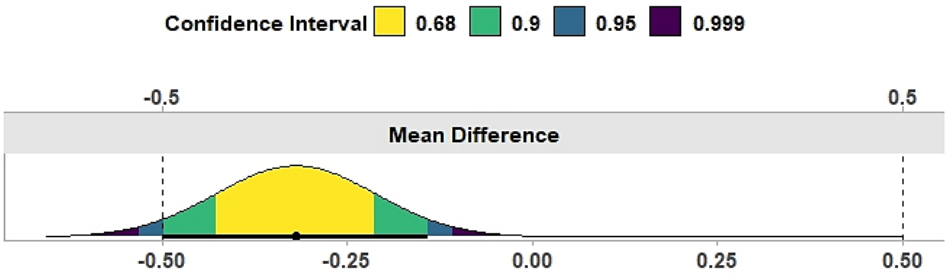


***Figura 9**. Visualizzazione degli intervalli di confidenza attorno alla differenza delle medie.*

È possibile dunque concludere che l'effetto sia statisticamente diverso da zero ma statisticamente troppo piccolo da essere considerato significativo per il nostro intervallo (Lakens, 2022).

Un test di equivalenza, quindi, se abbinato al test di significatività delle ipotesi nulle, permette di interpretare i risultati in modo più dettagliato grazie anche al fatto che viene esplicitamente richiesto di specificare il SESOI a priori. Ciò permette di trarre conclusioni che hanno una effettiva rilevanza pratica per quell'intervallo scelto, permette ad altri ricercatori di replicare il test e rende la ricerca falsificabile.

Se vengono applicati sia il test dell'ipotesi nulla sia il test di equivalenza si può giungere a quattro diversi risultati (***Figura 10***). L'effetto può essere statisticamente equivalente e non statisticamente diverso da zero (Situazione ***A***, ***Figura 10***). In questo caso si può concludere l'equivalenza a zero nell'intervallo. Un altro esito potrebbe essere statisticamente diverso da zero ma non statisticamente equivalente (***B***) e in questa circostanza nell'NHST si rifiuta l'ipotesi nulla ma non nel test di equivalenza. I casi ***C*** e ***D*** sono, infine, quelli analizzati sopra negli esempi. Il primo riguarda appunto un effetto statisticamente diverso da zero ma statisticamente equivalente a zero e in questo caso è possibile concludere l'equivalenza per quell'intervallo. Il secondo caso (***D***) avviene in particolare quando il campione risulta essere troppo piccolo per quello studio e infatti si avrà un effetto che non sarà né statisticamente equivalente a zero né statisticamente diverso (indeterminato) (Lakens, 2017).

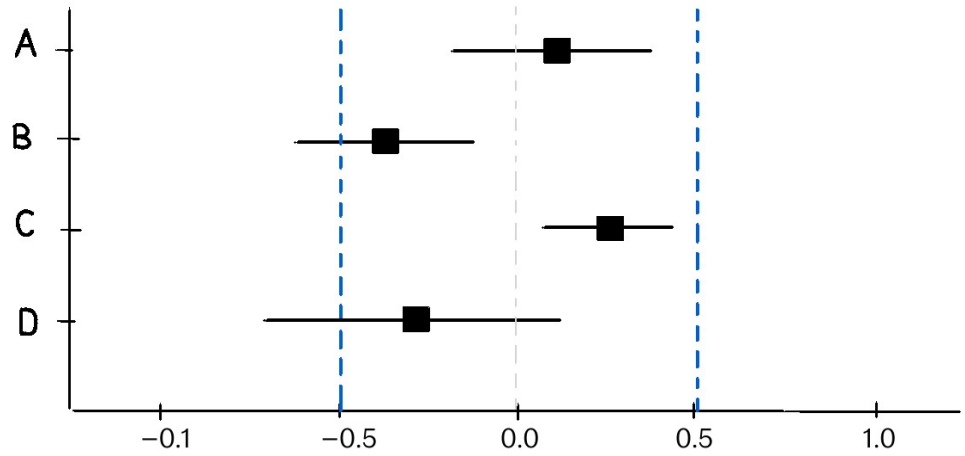


***Figura 10.*** *Differenze di medie con i loro rispettivi intervalli di confidenza nei limiti di equivalenza (-0.5 e 0.5) nei possibili esiti dei test di equivalenza.*

# 3. Determinare lo Small Effect Size Of Interest (SESOI)

Dal momento che non è possibile dimostrare che un effetto risulti esattamente uguale a zero, diventa fondamentale specificare quando quell'effetto è troppo piccolo per essere considerato interessante sia dal punto di vista teorico sia dal punto di vista pratico (Lakens, 2022). Bisogna quindi scegliere e giustificare un valore "soglia" entro il quale l'effetto può essere considerato equivalente a zero. Come già detto in precedenza, l'impostazione del SESOI è fondamentale e va fatta a priori. Un esperimento che non specifica un effetto minimo significativo non può essere confutato. Non solo è necessario per contestualizzare a fini pratici e statistici in che modo quello studio contribuirà alla conoscenza ma pone una base per discussioni future su quale sia l'effettivo effetto minimo di interesse in quel campo (Lakens et al., 2018). La giustificazione dei margini di equivalenza è strettamente legata all'importanza che lo studio avrà dal punto di vista clinico (Walker & Nowacki, 2011). Di conseguenza, specificare l'effetto minimo di interesse è la parte più critica e importante dei test di equivalenza perché determina l'esito dello studio ma gli dà anche credibilità scientifica (Walker & Nowacki, 2011). Inoltre, specificando l'effetto minimo di interesse si sottolinea la distinzione tra "statisticamente significativo" e "praticamente significativo", ossia da un punto di vista pratico, che sono concetti molto diversi ma che vengono spesso confusi (Anvari & Lakens, 2021).
Se tutti i ricercatori definissero un SESOI nei loro studi risulterebbe più semplice calcolare la potenza di un test nel rilevare effetti significativi, renderebbero le ricerche falsificabili (Lakens, 2022) e riproducibili.

## 3.1 SESOI stabiliti in base a criteri oggettivi

Nel caso in cui non vi fossero limiti teorici che impostano un effetto minimo di interesse che debba essere omogeneo per tutti allora un criterio oggettivo potrebbe riguardare sulla differenza minima percepibile (Lakens et al., 2018). Come descrivono Lakens e collaboratori in un articolo del 2018 (Lakens et al., 2018) un esempio potrebbe essere l'esperimento di Burriss et al. (2015) che era interessato a misurare il cambiamento del rossore sulle guance delle donne durante il periodo fertile. L'ipotesi era che gli uomini potessero percepire questo cambiamento ad occhio nudo e, nonostante fu smentita, siccome la differenza minima percepibile nel rossore del volto era qualcosa di misurabile fu possibile stabilire un SESOI oggettivo, ossia supportato da una teoria (Lakens, 2017).

## 3.2 SESOI stabiliti in base a criteri soggettivi.

Nel caso in cui non fosse possibile far utilizzo di giustificazioni oggettive, il SESOI dovrebbe essere basato sull'analisi costi-benefici (Lakens, 2022). Così facendo, risulta possibile valutare se un intervento ha un effetto abbastanza grande da essere considerato significativo in relazione al suo costo e alle risorse utilizzate. È chiaro che in questo caso sia i costi che i benefici siano soggettivi e quindi possono variare da ricercatore a ricercatore e nel tempo, ma questa analisi permette di valutare se un effetto è abbastanza grande da "meritare" le risorse necessarie per studiarlo in modo affidabile (Lakens et al., 2020).

Il SESOI può essere determinato anche dal campione massimo che si riesce a raccogliere per quello studio (Lakens et al., 2018). Le risorse e la quantità dei dati raccolti limitano l'esperimento e la possibilità di trarre conclusioni da esso (Lakens et al., 2018). Presupponiamo che, una volta impostato un alfa (α) di 0.05, il campione massimo che è possibile raccogliere sia di 100 partecipanti. Lo studio avrà una potenza statistica del 90% di rilevare un effetto di $d = 0.33$, allora i limiti di equivalenza saranno rispettivamente $\Delta_L = -0.33$ e $\Delta_U = 0.33$ (Lakens et al., 2018). In sostanza, nel momento in cui c'è un numero massimo di dati che viene raccolto, una volta impostato alfa, è possibile calcolare il SESOI basandosi sulla potenza statistica che quello studio può rilevare. È utile per porre una base per capire quali effetti possono essere rifiutati e quali invece meritano di essere studiati (Lakens et al., 2018) e in questo modo, si può dimostrare che l'effetto, se presente, non è abbastanza grande da poter essere rilevato su quel campione (Lakens, 2022).

Un altro modo per stabilire lo Small Effect Size Of Interest è basarsi su studi correlati. Nonostante non sia finora una pratica molto comune ritrovare specificato il SESOI negli studi presenti in letteratura (Lakens et al., 2018), negli studi di replicazione, o più in generale se si è interessati a testare la stessa ipotesi, è comunque possibile basarsi su lavori precedenti. Un metodo è lo *Small-Telescopes approach* introdotto da Simonsohn nel 2015 (Lakens et al., 2018). Propone di impostare il SESOI come l'effetto che avrebbe dato allo studio originale una potenza del 33% di rilevazione, assumendo che l'effetto esista. Anche in questo caso, quindi, si tratta di una giustificazione che parte dal numero del campione e, anche se non permette di concludere l'assenza di un effetto, può essere utilizzata come base per altri studi di replicazione perchè permette anche di valutare la dimensione del campione necessaria per ottenere risultati significativi.

Altri modi per impostare il SESOI basandosi su studi presenti concernono la media delle dimensioni dell'effetto riportate in letteratura e la dimensione più piccola che verrebbe ritenuta

significativa negli studi precedenti (Lakens et al., 2018). Questa giustificazione implicherà che il SESOI sia impostato sulla dimensione dell'effetto che lo studio originale aveva il 50% di potenza di rilevazione in un t-test indipendente. Questo approccio è in qualche modo simile allo Small Telescopes approach, sebbene porti a un SESOI leggermente più grande.

Un ulteriore metodo sono i "punti di riferimento" (*benchmarks*) per gli effetti, per esempio $d = 0.5$ (*Cohen's medium-sized effect*) (Lakens et al., 2018). Sono i meno preferibili perchè sono convenzioni arbitrarie che dovrebbero essere utilizzate solo in mancanza di altre raccomandazioni o basi chiare (Anvari & Lakens, 2021). Si tratta dunque della giustificazione più debole e quindi dovrebbe essere evitata (Lakens et al., 2018) perchè non si tratta di un effettivo indicatore di quale sia l'effetto minimo in quella ricerca.

## 3.3 Giustificare il SESOI: un esempio pratico per la misurazione delle esperienze soggettive.

È possibile stabilire la differenza minima rilevante utilizzando un "punto di ancoraggio (*anchor*) clinico" che riguarda valutazioni globali e relazioni cliniche del cambiamento avvenuto nel paziente dopo il trattamento (Anvari & Lakens, 2021). Nell'ambito psicologico, essendo la psicologia più orientata alle esperienze soggettive, è possibile utilizzare l'approccio dell'ancoraggio usando, come nei classici casi clinici, un giudizio globale ma basato su una valutazione globale soggettiva (autovalutazione del paziente) del cambiamento e utilizzando quest'ultima come punto di ancoraggio (Anvari & Lakens, 2021). Il metodo di valutazione globale quantifica, dunque, il più piccolo cambiamento ritenuto significativo a livello soggettivo. Si tratta di delineare quale sia la media dell'effetto sufficientemente grande da essere percepito, a livello soggettivo, come significativo. Questa differenza può essere utilizzata per giustificare il SESOI (Anvari & Lakens, 2021), perchè corrisponde ad un effetto minimo che ha una rilevanza pratica.

Viene chiesto al paziente il cambiamento percepito riguardo al costrutto di interesse in due momenti: T1, ossia prima dell'intervento e T2, dopo l'intervento. Gli item di ancoraggio (*anchor-item*) necessari sono uno per ogni dominio di ricerca e, una volta misurato il punteggio nei due momenti (T1 e T2), viene sottratto T1 da T2 per ogni partecipante. Successivamente questi verranno categorizzati in:

- individui che non hanno percepito cambiamento,
- individui che hanno percepito un minimo cambiamento,

- individui che hanno percepito un sostanziale cambiamento.

Calcolata la media per ogni categoria, la seconda (quelli che hanno percepito un cambiamento lieve) sarà quella presa in considerazione come stima della più piccola differenza soggettivamente percepita. In sostanza, la media di tali punteggi verrà utilizzata per stimare la differenza minima rilevante (Anvari & Lakens, 2021). È necessario, prima di utilizzare questo metodo di valutazione globale, che sia chiaro il perchè sia un metodo appropriato per la ricerca in questione, ossia possibilmente che si tratti di cambiamenti, esperienze e percezioni soggettive.

# 4. Applicazione di un test di equivalenza per studi di replicazione

Nell'ambito degli studi di replicazione è possibile applicare i test di equivalenza per valutare se lo studio originale è stato o meno replicato. In sostanza si va a valutare se la differenza tra l'effetto dello studio originale e l'effetto della replicazione sia abbastanza piccola da essere considerata equivalente a zero e, quindi, se i due studi sono uguali.

Il seguente studio di replicazione preso come esempio è tratto da Klein et al. (2023).

Bauer e collaboratori (2012) hanno esaminato se avere una mentalità consumistica riducesse la fiducia negli altri. L'ipotesi di partenza era che le persone avrebbero avuto una minore fiducia negli altri se avessero pensato a questi ultimi come a consumatori piuttosto che come individui. Hanno così proposto a 77 partecipanti una lettura in cui veniva presentato un dilemma sulla conservazione dell'acqua. Nella prima condizione il testo si riferiva ai lettori e alle altre persone nello scenario come "consumatori", mentre nella condizione di controllo veniva esclusivamente data l'etichetta di "individui". Successivamente, i partecipanti dovevano valutare quanto si fidassero del fatto che gli altri partecipanti avrebbero conservato l'acqua (scala da 1, per niente, a 7, molto). Dai risultati è emerso che nella condizione "consumatori" ($M$ = 4.08, $SD$ = 1.56) i partecipanti avevano meno fiducia nelle altre persone rispetto alla condizione di controllo ($M$ = 5.33, $SD$ = 1.30). Nello studio di replicazione, svolto su 6608 persone, sono stati ottenuti risultati simili per il gruppo sperimentale ($M$ = 3.92, $SD$ = 1.44) e più bassi per il gruppo di controllo ($M$ = 4.10, $SD$ = 1.45). Per essere sicuri che il test possa essere considerato replicato è possibile utilizzare un test di equivalenza per vedere se la differenza tra le medie dei due studi è abbasta piccola da essere considerata equivalente a zero.

Il primo passo da compiere è quello di impostare i limiti di equivalenza e cioè di giustificare un SESOI. Come abbiamo visto nel capitolo precedente, un metodo è lo *Small Telescopes approach*. Sarà necessario, dunque, calcolare l'effetto che avrebbe dato allo studio originale una potenza del 33% (***Figura 11***).

```
1. pwr::pwr.t.test( n = 77,
2.                  sig.level = 0.05,
3.                  power = 0.33,
4.                  type = "one.sample",
5.                  alternative = "two.sided" )

     One-sample t test power calculation

              n = 77
              d = 0.1753576
      sig.level = 0.05
          power = 0.33
    alternative = two.sided
```

***Figura 11.*** *Codice del pacchetto TOSTER (in blu) per calcolare, nei nostri dati, l'effetto che avrebbe dato allo studio originale (n = 77) una potenza del 33% (0.33) e output di Rstudio (in nero).*

Il SESOI sarà impostato quindi a 0.18.

A questo punto si applica il test di equivalenza (***Figura 12***).

```
1. TOSTER::tsum_TOST( m1 = 4.08,
2.                    sd1 = 1.56,
3.                    n1 = 77,
4.                    m2 = 3.92,
5.                    sd2 = 1.44,
6.                    n2 = 6608,
7.                    low_eqbound = -0.18,
8.                    high_eqbound = 0.18,
9.                    eqbound_type = "raw",
10.                   alpha = 0.05 )

Welch Modified Two-Sample t-Test

The equivalence test was non-significant, t(77.52) = -0.112, p = 4.56e-01
The null hypothesis test was non-significant, t(77.52) = 0.896, p = 3.73e-01
NHST: don't reject null significance hypothesis that the effect is equal to zero
TOST: don't reject null equivalence hypothesis

TOST Results
                t    df p.value
t-test     0.8956 77.52   0.373
TOST Lower 1.9031 77.52   0.030
TOST Upper -0.1119 77.52   0.456

Effect Sizes
             Estimate     SE             C.I. Conf. Level
Raw            0.1600 0.1787 [-0.1374, 0.4574]         0.9
Hedges's g(av)   0.1063 0.1201 [-0.0892, 0.3015]         0.9
Note: SMD confidence intervals are an approximation. See vignette("SMD_calcs").
```

***Figura 12****. Codice (blu) e output (nero) del test di equivalenza applicato per i nostri campioni, quello dello studio originale e della replica.*

La ***Figura 13*** mostra graficamente il risultato del test di equivalenza che risulta non significativo.

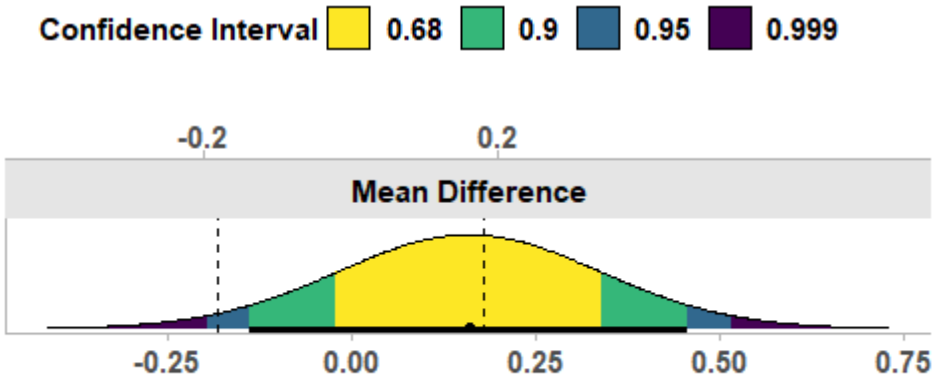


***Figura 13***. *Visualizzazione degli intervalli di confidenza attorno alla differenza delle medie*

Non è possibile concludere, dunque, né l'equivalenza né che ci sia un effetto statisticamente diverso da zero perché il risultato è inconcludente (indeterminato). Questo avviene, come ne abbiamo già discusso negli esempi precedenti, perché entrambi i test (NHST e test di equivalenza) risultano non significativi. Non ci sono dati sufficienti per rifiutare la presenza né di un effetto nel nostro intervallo né di un effetto più estremo.

In sostanza, per il nostro SESOI, non è possibile concludere che lo studio sia stato replicato perché i nostri risultati sono inconcludenti. L'esito del test di equivalenza è strettamente legato alla scelta dello Small Effect Size Of Interest ed è fondamentale giustificarlo. Nell'esempio appena presentato, se avessimo impostato un SESOI a 0.5, il test di equivalenza sarebbe risultato significativo e lo studio replicato. Così facendo, però si avrebbe a che fare con la giustificazione più debole, ossia tramite valori arbitrari (per esempio $d = 0.5$), che è quindi anche la meno condivisibile.

La criticità principale dei test di equivalenza è dunque la scelta e giustificazione dell'effetto minimo rilevante per il costrutto in questione. L'intervallo di equivalenza, infatti, essendo nella gran parte dei casi una scelta soggettiva, non per forza è condivisibile da tutti i ricercatori per questo diventa fondamentale che venga riportato. La conclusione di un test di equivalenza può essere condivisa solo se viene condivisa la giustificazione del SESOI (Lakens, 2022). Questo richiede un'attenta analisi dei dati che ci si aspetta di ottenere e della letteratura sull'argomento e non sempre è chiaro quale sia effettivamente l'intervallo oltre il quale l'effetto risulta essere clinicamente rilevante. A ciò si unisce un'ulteriore possibile problematica, ossia l'interpretazione dei risultati. Quest'ultimi vanno espressi in modo chiaro per evitare che siano soggetti a mal interpretazioni. Per esempio, se un test di equivalenza risulta significativo è

possibile rifiutare la presenza di un effetto più estremo del SESOI scelto ma non è possibile comunque concludere che l'effetto non esista, perché potrebbero esserci comunque effetti più piccoli (meno estremi) che non sono stati presi in considerazione. È possibile, dunque, stabilire che non ci sia un effetto solo specificando che ciò vale per quello specifico intervallo.

# Bibliografia


Anvari, F., & Lakens, D. (2021). Using anchor-based methods to determine the smallest effect size of interest. *J. Exp. Soc. Psychol*, 96:104159.

Conlquhoun, D. (2014). An investigation of the false discovery rate and the misinterpretation of p-values. *Soc. open sci.*, 1: 140216. 6.

*Dizionario di Filisofia* . (2009). Retrieved from Treccani: https://www.treccani.it/enciclopedia/teoria-della-falsificabilita_%28Dizionario-di-filosofia%29/

Ioannidis, P. A. (2005). Why most published research findings are false. *PLoS Med*, 2(8): e124.

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.

Klein, R. A., Vianello, M., Hasselman, F., Adams, B. G., Adams, R. B., Jr., Arper S., & Nosek, B. A. (2023). Many Labs 2: Investigating Variation in Replicability Across Sample and Setting. .

Lakens, D. (2017). Equivalence Tests: A Practical Primer for t Tests, Correlations, and Meta-Analyses. *Social Psychological and Personality Science*, 8(4), 355-362.

Lakens, D. (2019). The value of preregistration for psychological science: a conceptual analysis. *Japanese Psychological Review*, 62, 3, 221-230.

Lakens, D. (2022). *Improving Your Statistical Inferences.* Retrieved from https://lakens.github.io/statistical_inferences/.

Lakens, D., McLatchie, N., Isager, P. M., Scheel, A. M., & Dienes, Z. (2020). Improving Inferences About Null Effects With Bayes Factors and Equivalence Tests. *The Journals of Gerontology*, 75, 45–57.

Lakens, D., Scheel, A. M., & Isager, P. M. (2018). Equivalence Testing for Psychological Research: A Tutorial. *Advances in Methods and Practices in Psychological Science*, 1(2), 259-269.

Makel, M. C., Plucker, J. A., & Hegarty, B. (2012). Replications in Psychology Research: How Often Do They Really Occur? *Psychological Science*, 7(6), 537-542.

Nosek, A. B., Hardwicke, E. T., & Moshontz, H. (2022). Replicability, Robustness, and Reproducibility in Psychological Science. *Annual Review of Psychology*, 73, 1, 719–748.

O'Boyle, E., & Götz, M. (2022). Questionable Research Practices. In L. Jussim, J. A. Krosnick, & S. T. Stevens, *Research integrity: Best practices for the social and behavioral sciences* (pp. 260-294). Oxford University Press.

Otgaar, H., Riesthuis, P., Ramaekers JG, G. M., & L, K. (2022). The importance of the smallest effect size of interest in expert witness testimony on alcohol and memory. *Front. Psychol*, 13:980533.

Perezgonzalez, J. (2015). Fisher, Neyman-Pearson or NHST? A tutorial for teaching data testing. *Frontiers in Psychology*, 3;6:223.

Staddon, J. (2017). *Scientific Method: How Science Works, Fails to Work, and Pretends to Work (1st ed).* New York: Routledge.

Walker, E., & Nowacki, A. S. (2011). Understanding equivalence and noninferiority testing. *Gen Intern Med*, 26, 192–6.