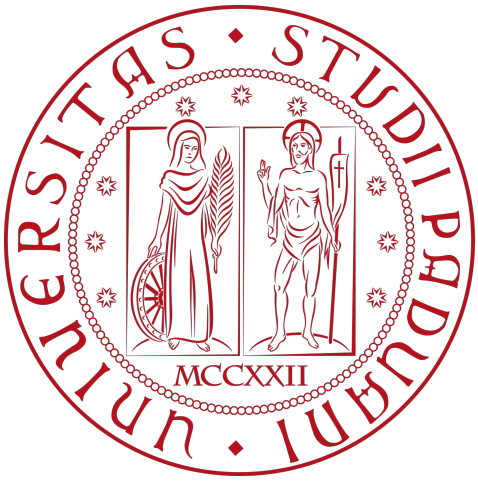
Università degli Studi di Padova
Dipartimento di Scienze Statistiche
Corso di Laurea Triennale in
Statistica per le Tecnologie e le Scienze



RELAZIONE FINALE

Test di NON-INFERIORITÀ

Relatore Prof. Laura Ventura
Dipartimento di Scienze

Laureanda: Sofia Lucia Catanese
Matricola N: 1217786

Anno Accademico 2022/2023

Indice

Capitolo 1: Introduzione

Nell'ambito della medicina vengono proposti nuovi trattamenti a malattie e il compito dello statistico è valutarli per testare se il loro effetto è superiore a quello del placebo o al trattamento standard/di riferimento già esistente.
Questo ambito di ricerca è discusso per la questione riguardante il test da utilizzare per valutare il nuovo farmaco proposto: a seconda della necessità vengono utilizzati diversi tipi di test o direzionali ad una coda e quindi test di non inferiorità o di superiorità, o test direzionali a due code [1].

Lo schema della relazione è il seguente. Nel capitolo 2 si introduce la differenza tra i test direzionali a una coda e nello specifico, il test di superiorità e il test non-inferiorità, per testare se c'è una differenza significativa tra un farmaco preesistente e/o il placebo e un farmaco nuovo proposto. Viene, poi, descritto il margine di non-inferiorità δ, essenziale per le analisi statistiche. Nel paragrafo 2.3, verranno presentati diversi approcci (intervalli di confidenza, test parametrici, test non parametrici) per risolvere l'ipotesi di non-inferiorità e il confronto tra essi. Infine, nel paragrafo 2.4, vengono discussi degli esempi e la sintassi utilizzata per fare le analisi statistiche, in particolare facendo riferimento al software R [2].

Capitolo 2:
Test Di Non-Inferiorità

Lo scopo di uno studio clinico è valutare se un trattamento sperimentale, per esempio un nuovo farmaco, non sia migliore solo in termini di efficacia o di frequenza/intensità degli effetti collaterali rispetto a farmaci standard o al placebo, ma che offra prestazioni migliori. Quindi, il focus centrale di una sperimentazione clinica sono il paziente e il suo stato di salute.

I cosiddetti test di non-inferiorità [3] appartengono alla grande famiglia dei test di equivalenza che si basano sul dimostrare l'equivalenza tra due trattamenti in studio. Si può dimostrare che una terapia è superiore a un trattamento di controllo, oppure si può dimostrare che la nuova terapia è non peggiore o equivalente di un trattamento efficace noto. In quest'ultimo caso, l'ipotesi verifica che le terapie sperimentali sono equivalenti alle terapie standard utilizzate definendo una soglia di differenza massima tollerabile, cioè è importante definire un intervallo compreso tra $-\delta$ e $+\delta$. In particolare, δ descrive il valore degli estremi dell'intervallo di confidenza predefinito tale per cui la differenza tra trattamenti sperimentali e quelli standard sia nulla, ossia che le due terapie siano equivalenti. In seguito, calcolato l'intervallo di confidenza, solitamente di livello 95%, della differenza tra i trattamenti sperimentali e quelli standard viene confrontato con l'intervallo di confidenza prescelto: l'equivalenza è vera se l'IC al 95% è contenuto nell'IC selezionato a priori. Si deve stare molto cauti quando si mostra con successo l'"equivalenza", cioè poca differenza tra un nuovo farmaco e un trattamento attivo, perché non dimostra di per sé che il nuovo trattamento sia efficace: "equivalenza" potrebbe significare che i trattamenti erano entrambi

efficaci nello studio, ma potrebbe anche significare che entrambi i trattamenti erano inefficaci nello studio.
Esistono poi il test di superiorità oppure il test di non-inferiorità, che sono detti anche test di equivalenza unilaterali. Il test di non-inferiorità ha come obiettivo di dimostrare che il nuovo trattamento non sia peggiore, entro un margine specificato, del trattamento standard utilizzato. Il nuovo trattamento è considerato “non-inferiore” rispetto all’altro se il limite inferiore di confidenza non è inferiore al valore prefissato per la differenza tra le due terapie. Il margine specificato, diverso caso per caso, è una scelta molto importante e intrigante per lo statistico dato che influisce sulla scelta statistica finale che si andrà a prendere per decidere la validità o meno della nuova terapia. In questo caso si definisce solamente l’estremo inferiore, $-\delta$, del margine di equivalenza. Di questo ne parleremo approfonditamente nel secondo paragrafo.
Questo modo di condurre le analisi va in contrasto con il test di superiorità, ossia un test per analizzare la superiorità di un trattamento rispetto a un altro: classico test unidirezionale che sta alla base dell’inferenza della statistica classica.
Ogni metodo può essere valido, ma ognuno richiede approcci inferenziali diversi. Uno studio ben progettato che mostra la superiorità di un trattamento rispetto a un controllo (placebo o terapia già preesistente) fornisce una forte evidenza dell'efficacia del nuovo trattamento, limitata solo dall'incertezza statistica del risultato. Infatti, è importante decidere la metodologia statistica su come affrontare lo studio e calcolare l’incertezza: verifica d’ipotesi piuttosto che intervalli di confidenza o, in alcuni casi, test non parametrici. Tutto ciò verrà affrontato nei seguenti paragrafi.

2.1 La struttura dell'esperimento clinico

L'ente responsabile della coordinazione e armonizzazione delle procedure in tutti i paesi dell'Unione Europea riguardanti le linee guida e indicazioni per la struttura migliore di un esperimento clinico è l'EMA [4]. In Italia, la sperimentazione clinica è posta sotto il controllo di autorità sanitarie come l'AIFA [5] e l'Istituto Superiore di Sanità che dettano precise disposizioni.

Ovviamente, una condizione necessaria tale per cui il nuovo prodotto venga commercializzato è il fatto che abbia un effetto nel trattamento e in particolare che si possa distinguere tra un effetto positivo e uno negativo. In presenza di un prodotto già in commercio per il trattamento della stessa patologia è necessario testare la non inferiorità del nuovo prodotto rispetto a quello già utilizzato e quindi utilizzare un test di non-inferiorità, oppure di superiorità a seconda dei singoli casi e dalla popolazione di riferimento.
In particolare, quando si parla di farmaci a confronto possiamo trovarci di fronte a due tipi di trial diversi: *placebo-controlled trial e active-controlled trial.*

2.1.1 *Placebo-controlled trial*

I *placebo-controlled trial* sono molto discussi tra gli studiosi, questo perché in questo tipo di studio i pazienti sono allocati casualmente a un gruppo di controllo che riceve un placebo e a un gruppo sperimentale, ossia a quale sarà somministrato il farmaco nuovo; oppure possono essere allocati in tre gruppi, uno sarà il gruppo a cui viene somministrato il placebo, uno il farmaco preesistente e uno il farmaco nuovo. Il problema nasce quando si parla di eticità: pazienti malati gravi di cui già esiste un trattamento standard per la loro malattia dovrebbero ricevere

quello come agente di controllo e non il placebo perché gli sarebbe negato un trattamento potenzialmente utile. Vedremo, in seguito, come uno studio “active-controlled” sia preferito rispetto a questo. Infatti, le linee guida federali affermano che prima di iniziare uno studio *placebo-controlled* anche clinicamente la differenza tra il placebo e il nuovo trattamento deve essere rilevante. Questo, è importante perché avere i dati relativi al placebo rende più efficienti gli esperimenti, dato che dimostra che il livello di efficacia di un nuovo trattamento potrebbe non essere sufficiente a giustificarne l’uso, e quindi richiede un numero di soggetti minore per determinare l’effetto del nuovo farmaco proposto.
Una volta confermata la fase clinica, si può procedere ad un’analisi statistica: confronto tra il trattamento di controllo e quello sperimentale. Questo tipo di analisi viene condotta attraverso un test bilaterale o unilaterale fissando un livello di significatività pari, rispettivamente a 0.05 e 0.025 oppure attraverso il calcolo di un intervallo di confidenza a una o due code, rispettivamente di livello 97.5% e 95%. Un test di non-inferiorità è solitamente valutato da un intervallo di confidenza bilaterale di livello 95%, che mostra un intervallo per la vera differenza tra il prodotto di prova e quello già esistente. La stima puntuale rappresenta la migliore stima della differenza reale, quindi se è positiva e se questa è tutta l'evidenza disponibile, è più probabile che il prodotto di prova sia migliore del riferimento e viceversa; mentre il limite inferiore dell'intervallo di confidenza viene solitamente interpretato come il grado di inferiorità rispetto al riferimento che può essere escluso sulla base dei dati presentati, anche se questo non è un limite inferiore effettivo perché l'entità dell'inferiorità potrebbe essere maggiore [6].

2.1.2 *Active-controlled trial*

I cosiddetti *active-controlled trial* sono studi clinici in cui non si ha la possibilità di avere dati relativi al placebo e quindi non si può valutare se

effettivamente il trattamento di prova proposto è superiore o inferiore a esso. In questi casi, infatti, si fa testo alla letteratura: se, per esempio, in un antecedente studio è stato provato un farmaco proposto per un certo tipo di trattamento, sotto certe condizioni quali numerosità campionaria, dose e altri parametri definiti, allora un *active-controlled trial* può essere ragionevole se riproduce l'impostazione in cui quel particolare farmaco è stato regolarmente efficace.
Questo tipo di ricerca risolve il problema della non eticità nell'utilizzare il placebo su pazienti malati e bisognosi del farmaco; però, riscontra anch'esso dei problemi: errore di secondo tipo alto, la possibile scarsa qualità dei metodi utilizzati nella ricerca e la non significatività della non differenza tra un nuovo farmaco e uno già efficace.

2.1.3 *Active-controlled* o *Placebo-controlled trial*?

Come si può leggere da Fisher e Kong (2001) attraverso un confronto randomizzato diretto (*active-controlled* trial) del trattamento sperimentale con il trattamento di controllo attivo, si deduce dai test di non inferiorità che il trattamento sperimentale sarebbe stato più efficace del placebo se un placebo fosse stato incluso nello studio [7].
Però, gli studi di non-inferiorità che non prevedono un confronto del trattamento con il placebo non posso essere utilizzati se per quella patologia non esiste già un trattamento di provata efficacia, o quando esiste una elevata risposta al placebo o una variabilità stagionale (per esempio nella demenza, nella depressione, ecc..).
Perciò se la domanda è quale dei due trial si preferisce, la risposta è che dipende dallo studio in questione [8].

È anche importante sottolineare che in alcuni campi, come l'oncologia e la pediatria, l'utilizzo del placebo per studi clinici non è assolutamente

etico: questo perché se non si hanno a disposizione dati utili per costruire un intervallo di confidenza per la differenza tra il trattamento nuovo e il placebo, si dovrà considerare il trattamento di controllo come placebo, e quindi condurre un test di superiorità del nuovo trattamento sul controllo.

2.2 La questione del margine di non-inferiorità

Un passo importante nella progettazione di uno studio di non inferiorità è specificare il margine di non-inferiorità δ (> 0). Questa è una questione molto delicata in quanto si va a scegliere un valore tale per cui un nuovo farmaco può essere validato o meno per la cura di una specifica malattia della persona; infatti, questa scelta deve sempre essere giustificata su basi sia cliniche che statistiche.

Di seguito verranno descritti il test di ipotesi da verificare e la procedura che si utilizzano per definire il margine di non-inferiorità e per finire verrà presentato un esempio della FDA (*Food and Drug Administration*) [9].

2.2.1 Test di ipotesi

In un *non-inferiority trial,* la non-inferiorità di un nuovo farmaco viene affermata se non si è a favore dell'ipotesi nulla (H_0) che il trattamento preesistente sia migliore di quello nuovo, decisione presa in base al margine di non-inferiorità. Quindi se accetto H_0 viene affermata la superiorità del farmaco già preesistente sul nuovo trattamento; mentre se accetto l'ipotesi alternativa (H_1) si afferma la non-inferiorità del nuovo trattamento rispetto a quello di controllo.
Il test di ipotesi che segue fa riferimento al valore della variabile di interesse misurata per il nuovo trattamento (T) a confronto con il valore per il trattamento di controllo (C):

$$\begin{cases} \mathrm{H}_0 \colon \mathrm{C} - \mathrm{T} \geq \delta \\ \mathrm{H}_1 \colon \mathrm{C} - \mathrm{T} < \delta \end{cases}$$

L'indirizzamento verso H_0 o H_1 si può risolvere con un test unilaterale sinistro a un livello di significatività α, oppure con un intervallo di confidenza simmetrico di livello $1 - \alpha$ per la differenza $C - T$, confrontando il limite superiore dell'intervallo di confidenza con il margine di non-inferiorità δ. L'ipotesi nulla quindi viene rifiutata, rispettivamente, se la statistica test è maggiore del quantile $1 - \alpha$ della distribuzione della variabile di interesse o se il limite superiore dell'intervallo è minore del margine.

2.2.2 Quale δ scegliere?

La definizione di margine di non-inferiorità è fondamentale e ovviamente l'EMA dà una linea guida sul come essere prudenti nello sceglierlo. In generale, la scelta di δ è basata su precisi obiettivi tenendo conto della letteratura e singolarmente di ogni studio clinico. Solitamente quest'ultimo rappresenta la minima differenza clinicamente accettabile, intendendo che il margine scelto non deve essere superiore alla più piccola differenza del risultato ottenuto quando si confrontano il farmaco di controllo e il placebo.

Il margine di non-inferiorità deve consentire di dedurre se il trattamento sperimentale è efficace oppure no sotto condizioni ben precise in riferimento alla sicurezza dei pazienti; infatti, per garantire che l'efficacia del farmaco nuovo proposto rispetto al placebo sia stabilita con grande fiducia è importante la conservazione di una "frazione" dell'effetto della terapia di controllo.

Le considerazioni statistiche sono attribuite al riassunto delle prove storiche del comparatore attivo e, quando possibile, al raggruppamento di una stima dell'effetto con un IC di livello 95% degli studi storici randomizzati controllati (per lo più controllati con placebo). Il margine δ è definito in base alla stima aggregata o in base al limite dell'IC che è il più vicino all'effetto nullo (in entrambe le situazioni, saranno chiamati M_1). Deve essere poi selezionata una frazione M_1 che è la frazione che deve essere conservata dal trattamento sperimentale, detta anche frazione conservata. Quindi, il margine rappresenta la frazione che rimane di M_1, chiamata M_2. Ad esempio, se si decide che il 75% di M_1 deve essere conservato dal nuovo farmaco per dimostrare la non-inferiorità, allora $M_2 = (1 - 0.75) \times M_1 = 0.25 \times M_1$ (o 25% di M_1).

L'interpretazione data da Snappinn e Jiang (2008) è che M_1 funziona come soglia di non-inferiorità e cioè che l'effetto del nuovo farmaco deve essere superiore alla frazione conservata M_1 e che la superiorità su un placebo è insufficiente. Solitamente M_1 è compresa tra il 50% *e l'*80%, ma per tipi di studi particolare come, per esempio, per testare gli antibiotici è del 90% [11].

Ci sono molti modi per calcolare δ. Nel seguito, verrà illustrato il metodo per calcolare il margine attraverso il rischio relativo: T, C *e* P denotano i tassi di incidenza di un evento clinico nella popolazione da cui sono stati campionati pazienti trattati, rispettivamente, al trattamento sperimentale, al trattamento di controllo e al placebo. Inoltre, C_0 e P_0 sono, i rapporti di incidenza relativi ai dati ricercati in letteratura, rispettivamente, per il trattamento di controllo e per il placebo. Considerando, quindi, l'effetto del trattamento in termini di rischio relativo il sistema di ipotesi mirato alla soluzione di test di non-inferiorità sarà il seguente:

$$\begin{cases} \mathrm{H}_0 : T/C \geq \delta \\ \mathrm{H}_1 : T/C < \delta \end{cases}$$

con T/C il rischio relativo di T rispetto a C [12].

Quindi a questo punto per risolvere il problema del valore del limite di non-inferiorità si può procedere in tre modi:

1. siccome δ deve rispecchiare una massima perdita di efficacia ammissibile legata al trattamento sperimentale rispetto al controllo, per consentire una perdita inferiore di $100 * (1 - \lambda)$ % dell'effetto di controllo da parte della terapia sperimentale in termini di rischio relativo ossia:
   $1 - T/P > \lambda * (1 - C/P)$ o, equivalentemente,
   $T/C < 1 + (1 - \lambda) * (P/C - 1)$,
   dove λ è una costante specifica compresa tra i valori 0 e 1 e
   $\delta = 1 + (1 - \lambda) * (P/C - 1)$;
2. alternativamente, si può considerare il rischio relativo su scala logaritmica. Lo schema di ipotesi diventa quindi:

$$\begin{cases} \mathrm{H}_0: \{\log(\mathrm{P}) - \log(\mathrm{T})\} / \{\log(\mathrm{P}) - \log(\mathrm{C})\} \leq \gamma \\ \mathrm{H}_1: \{\log(\mathrm{P}) - \log(\mathrm{T})\} / \{\log(\mathrm{P}) - \log(\mathrm{C})\} > \gamma \end{cases}$$

   sotto l'ipotesi H_1, se γ=0 allora il trattamento sperimentale è superiore al placebo in termini di efficacia, mentre, se $\gamma = 1$ allora il trattamento sperimentale è più efficace del controllo. Equivalentemente, si può utilizzare il seguente sistema di ipotesi:

$$\begin{cases} \mathrm{H}_0: \log(\mathrm{T}) - \log(\mathrm{C}) \geq \delta \\ \mathrm{H}_1: \log(\mathrm{T}) - \log(\mathrm{C}) < \delta \end{cases}$$

   con $\delta = (1 - \gamma) * \{\log(\mathrm{P}) - \log(\mathrm{C})\}$;

3. l'ultima possibilità è quella di attenersi alla letteratura considerando due valori ottimali e plausibili e scegliendo quello più piccolo.

Il margine di non-inferiorità δ è sconosciuto e stimabile, nelle migliori delle ipotesi, attraverso gli studi clinici passati e la letteratura; infatti, è

difficile impostare il sistema di verifica di ipotesi di non-inferiorità. Inoltre, qualsiasi dei tre metodi elencati sopra si usi bisogna tener conto dell'incertezza della stima dell'effetto di $(C - P)$ e spesso è più conveniente prendere come stima il più basso valore dei limiti inferiore per la quantità in questione forniti dai diversi studi clinici in letteratura [11].

2.2.3 Esempio *SPORTIF V*

Lo studio *SPORTIF V* (*Stroke Prevention Using Oral Thrombin Inhibitor in Atrial Fibrillation V*) è un esempio fornito nella guida della FDA sull'applicazione dei metodi a margine fisso e di sintesi. In particolare, viene studiata la non inferiorità di ximelagatran al warfarin in pazienti con fibrillazione atriale non valvolare per ridurre il rischio di complicanze tromboemboliche [10].
Il sistema di ipotesi di questo studio è il seguente:

$$\begin{cases} \mathrm{H}_0 : M_2 \geq (1 - 0.50) \times M_1 \\ \mathrm{H}_1 : M_2 < (1 - 0.50) \times M_1 \end{cases}$$

Ossia, è stato deciso che il ximelagatran dovrebbe preservare almeno il 50% dell'efficacia del warfarin, trovata negli studi passati, per essere considerata non inferiore. Il rischio relativo (RR) di warfarin rispetto al placebo nel calo di ictus ed embolia sistematica nei pazienti è di 0.36 con un intervallo di confidenza (IC) pari a $(0.25 - 0.53)$ con un livello del 95%. Quindi, come descritto in precedenza, si sceglie il limite superiore per definire il margine δ dato che è il limite più vicino all'effetto nullo. Questo limite rappresenta un rischio relativo di 1.90 del placebo rispetto al warfarin (aumento del 90% del rischio). Quindi, M_1 sarà 1.90 e M_2 1.38, quest'ultimo calcolando il 50% di M_1 su scala logaritmica e perciò il limite superiore dell'RR di ximelagatran rispetto al warfarin deve essere < 1.38. Però, il rischio relativo di ximelagatran rispetto a warfarin è stato di 1.39 con un IC pari a $(0.91 - 2.12)$ al 95%. Si conclude, quindi, dato che il

limite di superiore dell'intervallo di confidenza (2.12) supera il margine di non-inferiorità, che la non-inferiorità del ximelagatran rispetto al warfarin non è stata confermata.

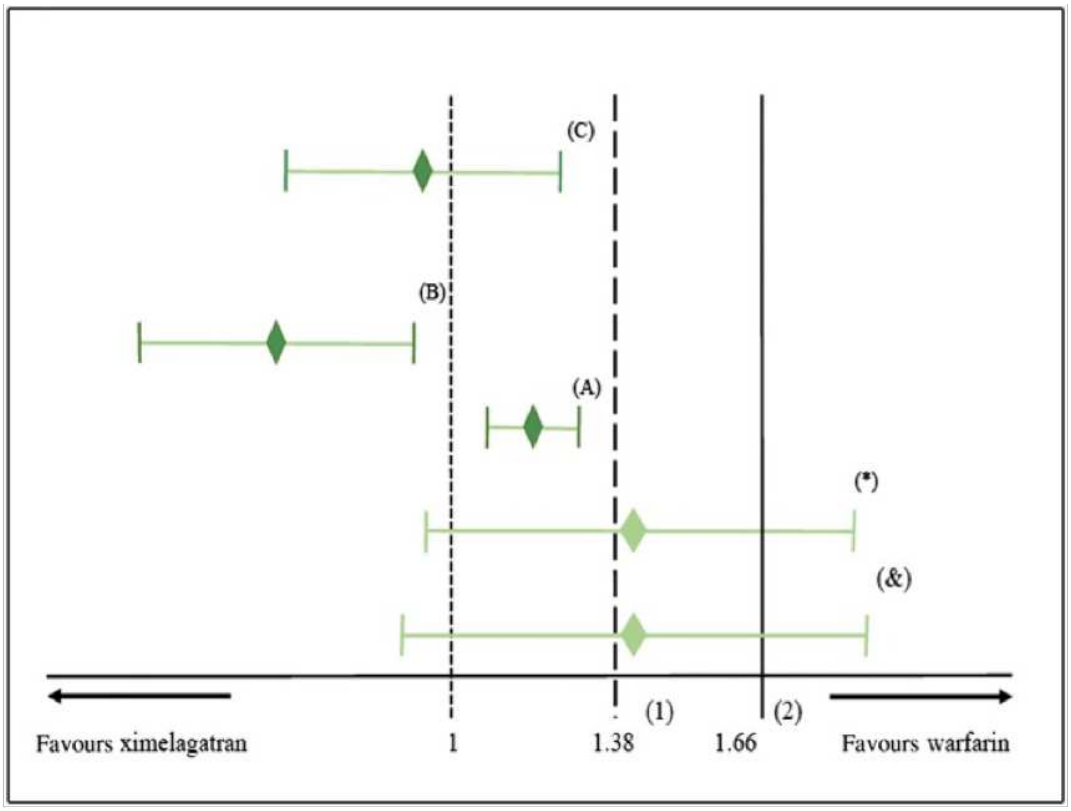


Figura 1: Intervalli di confidenza relativi al rischio relativo di ximelagatran rispetto a warfarin per testare la non-inferiorità

Nello studio poi viene applicato il metodo della stima e anche in questo caso la non-inferiorità non è provata.

In riferimento all'esempio appena descritto, la Figura 1 è utile per far interpretare meglio la situazione. In particolare, quelli contrassegnati con (1) e (2) sono i metodi applicati per verificare la non-inferiorità nello studio *SPORTIF V,* rispettivamente, il margine di non inferiorità calcolato con il margine fisso e con la stima puntuale (in questo ultimo caso, pari a 1.66, ossia il 50% del rischio relativo, pari a 2.77, su scala logaritmica). Inoltre, l'intervallo di confidenza originale dello studio *SPORTIF V* (*) è stato utilizzato per analizzare la non inferiorità con i metodi di margine fisso e di stima puntuale. In ogni caso, la non-inferiorità non è stata dimostrata perché il limite superiore dell'IC ha superato entrambi i margini

$(1.38\ e\ 1.66)$. Infine (&) risulta essere l'IC corretto del margine di non-inferiorità: anche in questo approccio la non-inferiorità non è verificata dato che il limite superiore dell'IC è maggiore del margine calcolato (1.66). Si conclude quindi, che la non-inferiorità sarebbe stata dimostrata esclusivamente se l'intervallo di confidenza si fosse trovato in una delle tre posizioni A, B o C.

2.3 Diversi approcci

In questa sezione verranno illustrati intervalli di confidenza e test per risolvere l'ipotesi di non-inferiorità. In particolare, in letteratura sono diversi gli approcci proposti: intervalli di confidenza e test con metodi parametrici, come l'approccio di Blackwelder [13,14], non parametrici, ossia il metodo TACT [15,16] e, infine, i test basati sui ranghi [17].

In generale, il metodo della statistica test Z_{pv} illustrato successivamente è noto per essere molto sensibile alla condizione di costanza nel tempo dell'effetto del trattamento di controllo. In particolare, se l'effetto del controllo attivo è peggiore nella popolazione a cui è stato somministrato il trattamento di controllo questo approccio non è valido nel senso che la probabilità di errore alfa supera il livello obiettivo, fissato convenzionalmente del 2.5%. Inoltre, l'approccio dell'intervallo di confidenza che utilizza la stima puntuale dell'effetto di controllo per definire il margine o il margine stimato δ* è peggiore o non migliore del metodo del test Z_{pv} in termini di probabilità di errore alfa. Al contrario, l'approccio dell'intervallo di confidenza che utilizza il limite inferiore dell'intervallo di confidenza del 95% dell'effetto di controllo attivo per definire il margine è molto meno sensibile alla condizione di costanza nel tempo. Infatti, sarà affrontato il metodo TACT come soluzione intermedia tra gli intervalli di confidenza e l'uso della statistica test. Infine, i test

basati sui ranghi vengono usati come alternativa dato che non è sempre possibile assumere che i dati in questione non soddisfano l'ipotesi di normalità.

2.3.1 Intervalli di confidenza

L'approccio dell'intervallo di confidenza è già stato in parte descritto con l'esempio dello studio *SPORTIF V.* L'intervallo di confidenza simmetrico viene costruito ad un livello $1 - 2\alpha$ per il $\log(T/C)$ e poi confrontato con il margine di non inferiorità δ: per rifiutare l'ipotesi nulla, ossia accettare l'ipotesi alternativa di non-inferiorità, l'IC deve essere inferiore a δ. Nel caso in cui il margine δ sia una costante fissa e nota, allora la probabilità di errore di I tipo α, ossia rifiutare H_0 erroneamente, sarà al massimo pari ad α. Mentre, nel caso in cui il margine di non-inferiorità sia stimato attraverso studi clinici simili passati, stima ottenuta solamente se la condizione di costanza è valida, fissato il valore di γ, allora nel calcolare l'errore di I tipo, bisogna tener conto anche dell'incertezza della stima di δ.

Distinguiamo ora i due casi per descrivere meglio le ipotesi [12].

1) Se nella valutazione dell'errore statistico si è tenuto conto dell'incertezza alla base della stima di δ, allora, scegliendo come stima di $\log(P) - \log(C)$ l'estremo inferiore dell'intervallo di confidenza di $\log(\widetilde{P_0}) - \log(\widetilde{C_0})$, allora gli estremi dell'intervallo di confidenza per $\log(P) - \log(C)$ sono definiti da:
   $[\log(\hat{\mathrm{T}}) - \log(\hat{\mathrm{C}}) \pm z_{1-\alpha/2} * \sigma_{TC}]$, e
   $\delta = (1-\gamma) * (log(\widetilde{P_0}) - log(\widetilde{C_0}) - z_{1-\alpha/2} * \sigma_{PC0})$,
   dove $\log(\hat{\mathrm{T}}) - \log(\hat{\mathrm{C}})$ è la stima di $\log(\mathrm{P}) - \log(\mathrm{C})$ del trattamento di controllo con il suo errore standard σ_{TC} e σ_{PC0} è lo standard error della quantità $\log(\widetilde{P_0}) - \log(\widetilde{C_0})$.

Perciò, si rifiuterà l'ipotesi nulla H_0, a favore dell'ipotesi alternativa di non-inferiorità, quando:

$\log(\hat{T}) - \log(\hat{C}) + z_{1-\alpha/2} * \sigma_{TC} < (1-\gamma)$ *[$(\log(\widetilde{P_0} - \log(\widetilde{C_0})) - z_{1-\alpha/2} * \sigma_{PC0}$].

Inoltre, la probabilità massima per la regione di rifiuto è

$\Phi(-z_{1-\alpha/2} * f)$, dove Φ è la funzione di ripartizione della distribuzione normale standard e

$f = \{\sigma_{TC} + (1-\gamma) * \sigma_{PC0}\} / \{\sigma_{TC}^2 + (1-\gamma)^2 * \sigma_{PC0}^2\}^{1/2}$, è una quantità sempre maggiore di 1. Quando la condizione di costanza nel tempo dell'effetto del farmaco di controllo è valida, la probabilità di errore α_1 dell'intervallo di confidenza che tiene conto dell'incertezza alla base della stima di δ, che utilizza il limite inferiore dell'IC per definire il margine di non-inferiorità, è sempre conservativa dato che vale sempre

$\alpha = \Phi(-z_{1-\alpha/2}) > \Phi(-z_{1-\alpha/2} * f) = \alpha_1$.

2) In un secondo caso, sempre se nella valutazione dell'errore statistico si è tenuto conto dell'incertezza alla base della stima di δ, allora, tenendo conto della stima puntuale di $\log(P) - \log(C)$, definita come $\log(\widetilde{P_0}) - \log(\widetilde{C_0})$, in questo caso si avrà evidenza contro l'ipotesi nulla H_0, quando:

$\log(\hat{T}) - \log(\hat{C}) + z_{1-\alpha/2} * \sigma_{TC} < (1-\gamma)$*$(\log(\widetilde{P_0}) - \log(\widetilde{C_0}))$.

Inoltre, la probabilità massima per la regione di rifiuto è

$\Phi(-z_{1-\alpha/2} * h)$, dove $h = \sigma_{TC} / \{\sigma_{TC}^2 + (1-\gamma)^2 * \sigma_{PC0}^2\}^{1/2}$, è una quantità sempre maggiore di 0, ma più piccola di 1. Quindi, la probabilità di errore α_1 dell'intervallo di confidenza della stima di δ, che utilizza la stima puntuale dell'effetto del trattamento di controllo per definire il margine di non-inferiorità, non è sempre conservativa. Infatti, l'effettivo valore di I tipo

$\alpha_1 = \Phi(z_{1-\alpha/2} * h) > \Phi(-z_{1-\alpha/2}) = \alpha$

e quindi la scelta meno conservativa potrebbe portare ad errori.

Dati i due problemi sopra riportati, potremmo chiederci se esiste un margine stimato δ^* tale che con l'approccio dell'intervallo di confidenza che utilizza questo margine abbia l'errore di I tipo α_1 (ossia il vero valore calcolato tenendo conto dell'incertezza che c'è alla base della stima di δ), esattamente uguale al livello desiderato e quindi corrispondente al valore di α. Si può dimostrare, quindi, che effettivamente esiste un certo margine di non-inferiorità che tenga conto di questa problematica e si calcola nel seguente modo:

$$\delta^* = -z_{1-\alpha/2} * \{[\sigma_{TC}^2 + (1-\gamma)^2 * \sigma_{PC0}^2]^{1/2} - \sigma_{TC}\} + (1-\gamma) * (\log(\widetilde{P_0}) - \log(\widetilde{C_0}))$$

che è una funzione esplicita dell'effetto del trattamento di controllo stimato e del suo errore standard σ_{PC0} ed anche funzione dello standard error σ_{TC} dell'effetto della terapia sperimentale rispetto a quella di controll o. La regione di rifiuto, ossia la regione per cui rifiuto l'ipotesi nulla H_0, è:

$$\log(\widehat{T}) - \log(\widehat{C}) + z_{1-\alpha/2} * \sigma_{TC} < \delta^*$$

che ha una probabilità di rifiutare l'ipotesi nulla quando quest'ultima è vera pari a:

$\alpha_1 = P\ [\ log(\widehat{T}) - \log(\widehat{C}) + z_{1-\alpha/2} * \sigma_{TC} < \delta^*] = P\ [\ log(\widehat{T}) - \log(\widehat{C}) - (1-\gamma)\ [\log(\widetilde{P_0}) - \log(\widetilde{C_0})]\ /\ \{[\sigma_{TC}^2 + (1-\gamma)^2 * \sigma_{PC0}^2]^{1/2} < -z_{1-\alpha/2}] = \Phi(-z_{1-\alpha/2}) = \alpha.$

Sorge ora il problema che l'errore standard σ_{TC} dipende anche dalla dimensione del campione e, conseguentemente, anche il margine stimato δ^* dipende dall'ampiezza del campione dello studio. Quest'ultima dovrebbe essere pianificata per testare la non inferiorità con un margine però, già prefissato con un certo errore di I tipo α ed errore di II tipo β. Ne consegue che chi progetta lo studio, per limitare le dimensioni del campione, tende a fissare margini quanto più ampi possibili. L'ampiezza dei margini influenza anche la probabilità di risultati

fuorvianti: più ampi sono i margini, maggiore è il rischio che un trattamento meno efficace risulti erroneamente non-inferiore; più ristretti sono i margini, più aumenta il rischio di rifiutare erroneamente un trattamento che è realmente non-inferiore.

2.3.2 Approccio delle statistiche test

Passiamo ora al metodo dell'utilizzo delle statistiche test. Questo criterio può essere utilizzato solamente nel momento in cui si può assumere la costanza nel tempo dell'effetto del trattamento di controllo sul placebo. In questo caso il test di ipotesi è:

$$\begin{cases} \mathrm{H}_0 : \log(\mathrm{T}) - \log(\mathrm{C}) \geq \delta \\ \mathrm{H}_1 : \log(\mathrm{T}) - \log(\mathrm{C}) < \delta \end{cases}$$

e la statistica test associata Z_{pv} è:

$$Z_{pv} = \log\left(\widehat{\mathrm{T}}\right) - \log\left(\widehat{\mathrm{C}}\right) - (1-\gamma) * [\log\left(\widetilde{\mathrm{P}_0}\right) - \log\left(\widetilde{\mathrm{C}_0}\right)] \,/\, \{[\sigma_{TC}^2 + (1-\gamma)^2 * \sigma_{PC0}^2]^{1/2}$$

Perciò si rifiuta H_0 se: $Z_{pv}^{oss} < -z_{1-\alpha/2}$

E quindi risulta che α_1 associato alla regione di rifiuto descritta sopra, è sempre uguale ad α [12].

Come detto prima, il metodo dell'utilizzo della statistica test è altamente sensibile alla condizione di costanza nel tempo. Infatti, se l'effetto della terapia di controllo è peggiore nella popolazione di controllo rispetto alla popolazione trovata in letteratura, questo approccio non è valido perché la probabilità di errore α_1 supera il livello di significatività α. Inoltre, il metodo dell'intervallo di confidenza che utilizza la stima puntuale dell'effetto del trattamento di controllo per definire il margine stimato δ^* da risultati meno attendibili in termini di probabilità dell'errore α rispetto all'approccio dell'IC che utilizza il limite inferiore dell'intervallo di

confidenza ad un livello $1 - \alpha$ dell'effetto della terapia di controllo per definire il margine δ di non-inferiorità che è molto meno sensibile al requisito di costanza nel tempo.

Nei seguenti paragrafi verranno presentati tre approcci differenti basati sul metodo della statistica test per testare la non-inferiorità, uno parametrico e due non parametrici, rispettivamente il metodo di Blackwelder chiamato "*The Blackwelder approach"*, il metodo TACT e il metodo basato sui ranghi.
Inoltre, si ricorda che si fa riferimento al valore della variabile di interesse misurata per il nuovo trattamento indicandolo con il simbolo T e al valore della variabile di interesse misurata per il trattamento di controllo con il simbolo C.

2.3.2.a *The Blackwelder approach* basato sulla differenza

In ambito clinico, per dimostrare che una terapia sperimentale è "*as least as good as*", ossia "almeno buona come", a una terapia standard, vengono usati test o intervalli di confidenza che devono escludere l'inferiorità clinica, della terapia sperimentale, con alta probabilità. Questo approccio è definito dall'unione di quattro stadi di natura:

1. Inferiorità clinica;
2. Tolleranza clinica;
3. Equivalenza;
4. Superiorità.

Perciò, Blackwelder ha proposto un test unilaterale per rigettare l'ipotesi nulla che il trattamento standard sia "migliore" del trattamento sperimentale, a favore dell'ipotesi alternativa che mostra il criterio "*as least as good as*" [13,14].

In particolare, in questo paragrafo verrà presentato l'approccio di Blackwelder basato sulla differenza delle medie relative alla variabile di interesse quando essa è misurata per il nuovo trattamento (μ_T) e quando, invece è misurata per il trattamento di controllo (μ_C).
In riferimento al seguente schema di ipotesi,

$$\begin{cases} \mathrm{H}_0 : \mu_C - \mu_T \geq \delta \\ \mathrm{H}_1 : \mu_C - \mu_T < \delta \end{cases}$$

con $\delta > 0$ margine di non-inferiorità, possiamo chiarire le regioni di rifiuto del test corrispondenti ai quattro stadi di natura:

1. Inferiorità clinica: $\mu_C - \mu_T \geq \delta$, quindi vuol dire che la terapia sperimentale è inferiore di una quantità del valore di δ (o più), rispetto alla terapia di controllo;
2. Tolleranza clinica: $0 < \mu_C - \mu_T < \delta$, perciò significa che la terapia sperimentale è inferiore di una quantità al più di δ rispetto alla terapia di controllo;
3. Equivalenza: $\mu_C - \mu_T = 0$, che rappresenta che le due terapie sono equivalenti;
4. Superiorità: $\mu_C - \mu_T < 0$, quindi indica che la terapia sperimentale è superiore rispetto alla terapia di controllo.

2.3.2.b *The Blackwelder approach* basato sul rapporto

Ora, invece, si considera la relazione tra le due medie, μ_C e μ_T, come un rapporto: $R_{true} = \mu_T/\mu_C$, con $\mu_C \neq 0$ che corrisponde al vero valore del rapporto tra le due medie.
Quindi, il sistema di ipotesi per la verifica della non-inferiorità sarà:

$$\begin{cases} \mathrm{H}_0 : \mu_T/\mu_C \geq R \\ \mathrm{H}_1 : \mu_T/\mu_C < R \end{cases}$$

con $R < 1$, limite inferiore calcolato sulla base di una percentuale che indica una soglia per la tolleranza clinica (solitamente compresa tra l'80% e il 90%).
Ancora una volta, per mostrare '*as least as good as*', si rifiuta H_0 a favore di H_1 [13,14].
Per quanto riguarda le interpretazioni delle ipotesi in relazione ai quattro stadi di natura, in questo caso, si ha:

1. Inferiorità clinica: $\mu_T/\mu_C \geq R$, quindi vuol dire che la terapia sperimentale ha un'efficacia inferiore del $(R * 100)\%$, rispetto all'efficacia della terapia di controllo;
2. Tolleranza clinica: $R < \mu_T/\mu_C < 1$, perciò significa che la terapia sperimentale ha una perdita di efficacia al più del $((1 - R) * 100)\%$ rispetto alla terapia di controllo;
3. Equivalenza: $\mu_T/\mu_C = 1$, che rappresenta che le due terapie sono equivalenti;
4. Superiorità: $\mu_T/\mu_C > 1$, quindi indica che la terapia sperimentale è superiore rispetto alla terapia di controllo.

Quindi, utilizzare questo tipo di approccio, rispetto all'approccio della differenza tra medie, ha il vantaggio che la terapia sperimentale può essere vista come percentuale della risposta della terapia standard, però i due test sono equivalenti se si guarda dal punto di vista dell'interpretazione del rapporto esistente tra i due trattamenti. Infatti, i due metodi vengono a coincidere quando nel test di Blackwelder si scelga come valore per δ, una piccola percentuale $(1 - R)\,\%$ della media μ_C del trattamento di controllo.
Allora, in questo caso si ha:

$$\delta = (1 - R) * \mu_C,$$

e quindi l'ipotesi H_0: $\mu_C - \mu_T \geq \delta$,

con i seguenti passaggi

$$\mu_C - (1 - R) * \mu_C \geq \mu_T$$
$$\mu_C - \mu_C + R * \mu_C \geq \mu_T$$
$$R * \mu_C \geq \mu_T$$
$$R \geq \mu_T / \mu_C ,$$

che è l'ipotesi nulla H_0 del test di Blackwelder basato sul rapporto tra medie.

2.3.2.c Test e intervalli di confidenza per l'approccio di Blackwelder

La verifica di ipotesi tramite l'approccio del rapporto tra medie di Blackwelder, si può affrontare in due modalità: o tramite l'uso di una statistica test unidirezionale da confrontare con il valore critico ad un livello di significatività $1 - 2\alpha$ scelto a priori, o attraverso la costruzione di un intervallo di confidenza simmetrico per il vero valore del rapporto tra le due medie chiamato R_{true} da confrontare con il valore prescelto R [14].

Considereremo come stimatore per R_{true}, $\hat{R} = \bar{X}_T / \bar{X}_C$, che è asintoticamente non distorto e normalmente distribuito. In particolare, si presuppone che $\bar{X}_T \sim$ N (μ_T,σ_T^2) e $\bar{X}_C \sim$ N (μ_C,σ_C^2); che le varianze siano tutte uguali, ossia $\sigma_T^2 = \sigma_C^2 = \sigma$; che la numerosità campionaria sia uguale, ossia $n = n_T = n_C$.

Per quanto riguarda il test di ipotesi, bisogna applicare la seguente riparametrizzazione:

$$\begin{cases} \mathrm{H}_0 \colon \mu_T - R * \mu_C \leq 0 \\ \mathrm{H}_1 \colon \mu_T - R * \mu_C > 0 \end{cases}$$

ossia, si considera come decisore tra l'inferiorità e la non-inferiorità del trattamento sperimentale su quello di controllo il segno della differenza

esistente tra la media della variabile indicatrice dell'efficacia del farmaco sperimentale e la percentuale della media del farmaco di controllo. Infatti, la statistica test associata alla verifica di ipotesi risulta avere una distribuzione tale da permettere di costruire intervalli di confidenza e test e si costruisce un test uniformemente più potente:

$T_{test} = (\bar{X}_T - R * \bar{X}_C) \sim t_{2*n-2}$, ossia la statistica test T_{test} si distribuisce come una *t di student* con gradi di libertà $2n - 2$.

Mentre, per quanto riguarda gli intervalli di confidenza simmetrici, centrati in $\hat{R}$, ad un livello $1 - 2\alpha$, l'intervallo di confidenza per R_{true} risulta essere, approssimativamente:

$$\hat{R} \pm z_\alpha [s^2 \ * \ (1 - R^2) / \ n * (\bar{X}_C)^2]^{1/2}$$

dove s^2 è la varianza corretta stimata.
Questo intervallo è basato su

$$(\hat{R} - R) / \ [s^2 \ * \ (1 - R^2) / \ n * (\bar{X}_C)^2]^{1/2}$$

che non dovrebbe superare, in valore assoluto, la quantità $t_{\alpha;2*n-2}$. I valori di R dove il rapporto è effettivamente uguale alla costante critica appena citata sono le radici di una complessa equazione quadratica. Tuttavia, per campioni di grandi dimensioni, le radici sono approssimativamente ai valori che forniscono un intervallo di confidenza approssimato per R_{true} riportato prima.
Per decidere se rifiutare o meno l'ipotesi nulla a favore di quella alternativa di non-inferiorità, si deve controllare che il valore di R sia interamente all'interno dell'intervallo di confidenza, ma soprattutto che contenga solo valori per R_{true} più grandi di R.

Ricapitolando, l'intervallo di confidenza è centrato in $\hat{R}$ e deve essere osservato un valore più grande di R se si vuole costruire un intervallo che

abbia qualche possibilità di escludere il valore R e perciò di rifiutare l'ipotesi nulla a favore di quella di non-inferiorità. Quindi, lo standard error associato a $\hat{R}$ sarà più grande quando si usa l'approccio dell'intervallo di confidenza rispetto al test standard e di conseguenza si ha che metodo dell'IC potrebbe portare ad un valore del limite inferiore non affidabile. Perciò l'uso del test standard risulta essere più efficiente rispetto al metodo dell'intervallo di confidenza.
In particolare, la dimensione campionaria ottimale per il test standard, ad un livello di significatività pari ad α e una potenza del test pari a β, risulta essere la seguente: $n = [(CV)^2 * (z_{1-\alpha} - z_{\beta})^2(1 + R^2)]/(R_{TRUE} - R)^2$.

2.3.2.d Approcci non parametrici

È possibile applicare anche metodi non parametrici quando non è possibile affermare la normalità della distribuzione dei dati in esame, per questo motivo si propongono il metodo il metodo TACT e il metodo dei ranghi per studiare la non-inferiorità.

TACT (*Two-stage active control testing*)

Abbiamo già accennato che, sotto l'assunzione di costanza nel tempo dell'effetto del farmaco di controllo, è più efficiente utilizzare una statistica test per individuare la regione di accettazione e di rifiuto dell'ipotesi di non-inferiorità, rispetto agli intervalli di confidenza. Ma se non ci sono i presupposti, questo comporta un aumento dell'errore di I tipo α, mentre questo non accade per gli intervalli di confidenza che sono un metodo più robusto in relazione a questo aspetto. Il metodo TACT, quindi, è stato sviluppato come valido compromesso tra i due approcci [15,16].

Si vuole testare l'efficacia di un farmaco sperimentale in termini di efficienza di un determinato evento: di seguito verranno specificate le procedure del metodo illustrato.
Verranno adottate le stesse notazioni utilizzate in precedenza per specificare gli effetti dei vari trattamenti nell'esperimento in corso e relative ai dati storici $(T, C, P, C_0 e\ P_0)$.
Considerando valida l'assunzione di costanza nel tempo dell'effetto del trattamento di controllo e assumendo che l'effetto del placebo sia immutato rispetto a quello riscontrabile nei dati storici, si può scrivere, su scala logaritmica per il rischio relativo, il seguente modello per il nuovo effetto del trattamento:

$$\log(event\ rate) = \mu\ + \beta * X_{C_0} + \gamma * X_C + \zeta * X_T,$$

con $\mu = \log P = \log P_0$ effetto del placebo, $\beta = \log C_0/P$ e $\gamma = \log C/P$ che denotano, rispettivamente l'effetto della terapia di controllo sul placebo nella popolazione dei dati storici e dell'evento in studio, $\zeta = \log T/P$ l'effetto della terapia sperimentale. Inoltre, $X_{C_0}, X_C\ e\ X_T$ sono le variabili indicatrici associate ai vari trattamenti.
Quindi, l'ipotesi di efficacia del trattamento sperimentale può essere la seguente:

$$\begin{cases} \mathrm{H}_0: \zeta \geq 0 \\ \mathrm{H}_1: \zeta < 0. \end{cases}$$

Mentre, l'ipotesi del test di non inferiorità, ossia la capacità della terapia sperimentale di preservare il $100 * \lambda\%$ dell'effetto della terapia di controllo, è:

$$\begin{cases} \mathrm{H}_0: \zeta - \gamma \geq \delta \\ \mathrm{H}_1: \zeta - \gamma < \delta, \end{cases}$$

con $\delta = -(1-\lambda) * \gamma$ la percentuale dell'effetto di controllo che vogliamo che il nuovo farmaco conservi.

Il metodo TACT consiste in due passaggi:

1. Deve essere dimostrata la superiorità del trattamento di controllo sul placebo dai dati storici: se questo viene dimostrato allora si può procedere con la fase successiva;
2. La fase successiva consiste in due fasi di analisi, che coinvolgono sia le prove storiche che lo studio in esame:
   I. Se l'effetto di controllo è molto inferiore nella popolazione in studio rispetto alle popolazioni dei dati storici, ossia se $\gamma \gg \beta$, allora sia il metodo degli intervalli di confidenza sia quello basato sui test standard possono portare a un errore di I tipo α grande. Quindi, in questo primo stadio si verifica, attraverso una statistica test, la costanza nel tempo, che serve per capire se fermare o meno lo studio in questione al tempo t $(0 < t < 1)$. In particolare, la statistica test deve stimare la differenza tra gli effetti registrati dal trattamento di controllo nei diversi tempi, dati storici ed evento in studio:

$$Z_{C_t} = \frac{\log \hat{C}_t - \log \hat{C}_0}{SE},$$

   dove $\hat{C}_t$ e $\hat{C}_0$ sono i tassi degli eventi stimati del trattamento di controllo, rispettivamente, al tempo t e al tempo $t = 0$ e SE è l'errore standard relativo alla quantità al numeratore. La statistica test in questione deve essere confrontata con un certo valore fissato U_t: se $Z_{C_t} > U_t$ allora si rifiuta l'ipotesi di costanza nel tempo dell'effetto della terapia di controllo e perciò è consigliato sospendere lo studio di non-inferiorità in esame;
   II. Si può procedere con lo stadio successivo se l'ipotesi di costanza nel tempo è valida, ossia se viene rifiutata l'ipotesi che $\gamma \gg \beta$. In particolare, ora si considera la statistica test al tempo $t = 1$, ossia Z_{C_1}:

- se $Z_{C_1} > U$ allora bisogna fermare le analisi di non-inferiorità perché potrebbe essere plausibile che $\gamma \gg \beta$;
- se, invece, $Z_{C_1} < L$, allora si può decidere se affrontare l'ipotesi di non-inferiorità con il metodo standard se la statistica test è particolarmente ridotta, oppure il metodo degli intervalli di confidenza se il valore di Z_{C_1} esprime una certa differenza di effetto della terapia di controllo.

Il metodo del test standard per il test di non-inferiorità usa la seguente statistica test:

$$Z_{pv} = \frac{\log(\hat{T}) - \log(\hat{C}) - (1-\lambda) * (\log(\widetilde{P_0}) - \log(\widetilde{C_0}))}{\sqrt{\sigma_{TC}^2 + (1-\gamma)^2 * \sigma_{PC0}^2}},$$

con $\log(\widehat{\mathrm{T}}) - \log(\hat{\mathrm{C}})$ stima di $log\, T - log\, C$ con standard error σ_{TC} e σ_{PC0} standard error di $\log(\widetilde{\mathrm{P_0}}) - \log(\widetilde{\mathrm{C_0}})$.
Perciò si rifiuta l'ipotesi nulla se $Z_{pv} < -z_{1-\alpha}$.
Mentre, la regola per il rifiuto dell'ipotesi nulla con il metodo dell'intervallo di confidenza per la differenza risulta essere:
$\log(\hat{T}) - \log(\hat{C}) + z_{1-\alpha} * \sigma_{TC} < (1-\gamma) * (\log(\widetilde{P_0}) - \log(\widetilde{C_0}) - z_{1-\alpha} * \sigma_{PC0})$.
In particolare, i limiti L e U sono espressi come multipli della deviazione standard di Z_{C_1} e possono essere calcolati tramite simulazione. Ad esempio, per verificare H_0 una scelta ragionevole è $L = z_{1-\alpha} = 1.645$, con $\alpha = 0.95$ e $U = 3$.

Sono stati eseguiti numerosi studi di simulazione Monte Carlo per esplorare l'errore α e le prestazioni della potenza $(1-\beta)$ del metodo TACT, in particolare per malattie cardiovascolari, dato che, solitamente, in tali studi i tassi di incidenza nella popolazione da cui sono stati campionati i pazienti trattati con il trattamento di controllo (C) e con il placebo (P) hanno valori bassi.

Metodo basato sui ranghi

Le statistiche test basate sui ranghi hanno il vantaggio di essere semplici da trattare e di essere più robuste delle statistiche test che utilizzano direttamente le osservazioni numeriche campionarie.

Consideriamo più centri clinici, supponiamo che quelli da inserire nel campione siano stati scelti in modo casuale e che la somministrazione del farmaco di controllo, di quello sperimentale e, eventualmente, del placebo, ai pazienti sia stata fatta anch'essa in modo casuale. Per applicare il metodo dei ranghi le assunzioni di base sono:

1. la numerosità interna a ciascun centro deve essere piuttosto elevata (> 30);
2. i centri sono scelti in modo casuale;
3. ci sia un solo rilevatore per ogni centro in modo da garantire l'omogeneità delle misurazioni.

Allora, in generale [17]:

- Consideriamo *t* trattamenti, *c* centri clinici e n_{ij} replicazioni dell'*i-esimo* trattamento nel *j-esimo* centro, con $i = 1, .., t$ e $j = 1, .., c$;
- Y_{ijl} è la singola osservazione della *l-esima* replicazione dell'*i-esimo* trattamento nel *j-esimo* centro, con $l = 1, \dots, n_{ij}$.

Quindi un modello generale, adeguato a studi clinici è:

$$Y_{ijl} = \theta_i + \beta_j + e_{ijl}$$

dove:

i. θ_i è l'effetto medio dell'*i-esimo* trattamento;
ii. β_j è l'effetto del *j-esimo* centro e si assume che esse siano variabili indipendenti e identicamente distribuite (i.i.d.) di media 0 e di varianza σ_β^2;
iii. e_{ijl} è l'errore casuale si assume che anche esse siano variabili i.i.d. di media 0 e varianza σ_e^2;
iv. $\beta_j \; e \; e_{ijl}$ sono indipendenti.

Sotto l'assunzione che l'effetto del trattamento sia simile nei diversi centri si considera la nuova variabile $\varepsilon_{ijl} = \beta_j + e_{ijl}$ e l'interesse è $\theta_i - \theta_{i^*}$ con $i \neq i^* = 1, \ldots, t$.

Quindi si considera il vettore di dimensione $n_{\cdot j} \times 1$ delle singole osservazioni $Y_j = (y_{1j}, \ldots, y_{ij}, \ldots, y_{tj})^T$ per il *j-esimo* centro con $y_{ij} = (y_{ij1}, \ldots, y_{ijn_{ij}})$ vettore delle singole osservazione del gruppo trattato con l'*i-esimo* trattamento nel *j-esimo* centro clinico e quindi ε_j sarà il vettore degli errori del modello considerato per il centro clinico *j*.

Nell'inferenza parametrica si ha che la ε_j sono variabili casuali normali multivariate a media 0 e matrice di varianze e covarianze pari a $\sigma^2 * [(1-\rho) * I_{n_{\cdot j} \times n_{\cdot j}} + \rho * \mathbf{1}_{\boldsymbol{n_{\cdot j}}} * \mathbf{1}'_{\boldsymbol{n_{\cdot j}}}]$, con $\sigma^2 = var(\varepsilon_{ijl}) = \sigma_\beta^2 + \sigma_e^2$,

$\rho = {\sigma_\beta^2}\Big/{(\sigma_\beta^2 + \sigma_e^2)}$ coefficiente di correlazione tra due componenti di ε_j, $I_{n_{\cdot j} \times n_{\cdot j}}$ matrice identità di ordine $n_{\cdot j}$ e $\mathbf{1}_{\boldsymbol{n_{\cdot j}}}$ vettore unitario di dimensione $n_{\cdot j} \times 1$.

Nelle applicazioni l'assunzione della distribuzione del vettore degli errori normale multivariata è spesso violata e quindi si ha la necessità di adottare altri metodi. Quella illustrata di seguito è una metodologia basata sulla minimizzazione della somma delle funzioni di dispersione di Jaeckel, detta *Jaeckel-type dispersion function,* basata sui ranghi dei residui del modello assunto per ottenere la stima di θ_i.

La funzione di dispersione di *Jaeckel* (1972) basata sui ranghi dei residui per il *j-esimo* centro è utilizzata per fare inferenza per studi clinici dove gli n_{ij} sono grandi e il numero di centri clinici *c* è fissato, con $j = 1, \ldots c$. Indicato $\theta = (\theta_1, \ldots, \theta_t)^T$ si ha che la funzione, basata sui punteggi di *Wilcoxon*, è [17]:

$$D_j(\theta) = \sqrt{12} \sum_{i=1}^{t} \sum_{l=1}^{n_{ij}} a\big(W_{ijl}\big) * W_{ijl} \,,$$

che è una funzione lineare a tratti dei residui e quindi l'influenza rispetto ai valori anomali nella stima finale è minore, continua, non negativa, convessa in θ, con $a(W_{ijl}) = (n_{.j} + 1)^{-1} * R(W_{ijl}) - 1/2$, $W_{ijl} = Y_{ijl} - \theta_i$ e $R(W_{ijl})$ corrisponde al rango del residuo W_{ijl} del trattamento *i* nel centro *j*. Da notare che i centri clinici più grandi non domineranno su quelli più piccoli in quanto i ranghi dei residui $D_j(\theta)$ sono pesati con $(n_{.j} + 1)^{-1}$ e il coefficiente dei residui è $W_{ijl} = Y_{ijl} - \theta_i \pm \mathbf{1/2}$, con $i = 1, \dots, t$ e $j = 1, .., c$.

Quindi la funzione combinata di dispersione è definita nel seguente modo:

$$D(\theta) = \sum_{j=1}^{c} D_j(\theta) ,$$

che è anch'essa una funzione lineare dei residui, continua, non negativa e convessa in θ. Si nota, inoltre, che per $t = 2$, $D(\theta)$ si riduce a una funzione a un solo parametro: si considera $\theta_1 = \mu + \delta$ e $\theta_2 = \mu$, quindi si ha che $D(\theta_1, \theta_2) = D(\mu + \delta, \mu) = \sum_{j=1}^{c} D(\mu + \delta, \mu) = \sum_{j=1}^{c} D_j(\delta, 0) = D(\delta, 0)$, dove $\delta = \theta_1 - \theta_2$.

Nella Figura 2 è riportato il grafico di $D(\delta, 0)$ in funzione di shift $\delta = \theta_1 - \theta_2$, quando $c = 2$, $n_{ij} = 4$, $\rho = 0.5$, $\sigma = 1$ e i dati sono generati da una *t di Student* con 3 gradi di libertà.

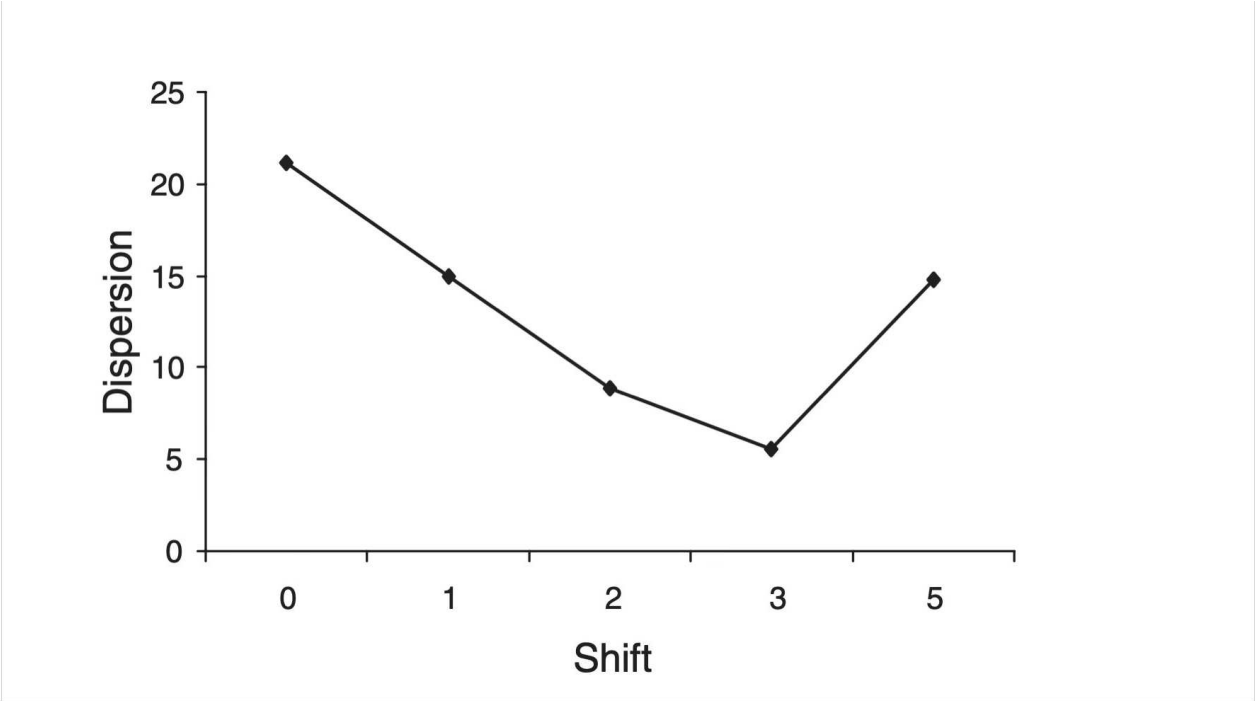


Figura 2 : Grafico della funzione di dispersione quando i dati sono generati da una t di student

Si nota che $D(\delta, 0)$ raggiunge il minimo in 3.
Quindi, lo stimatore $\hat{\theta}$ basato sui ranghi per il vettore degli effetti dei trattamenti θ si ottiene minimizzando la funzione di dispersione $D(\theta)$. Di conseguenza, quindi, per il caso descritto sopra si ottiene anche la stima di $\hat{\delta}$.
Nel caso specifico di interesse con due trattamenti, uno di controllo e uno sperimentale, somministrati a due gruppi di pazienti $(t = 2)$, all'interno di un unico centro $(c = 1)$, si ha che lo stimatore $\hat{\delta}$, corrisponde allo stimatore di *Hodges-Lehmann* (2006) di $\delta = \theta_1 - \theta_2$ basato sul test dei ranghi di *Mann-Whitney* [17].

Quando si parla di studi di non-inferiorità, il problema può essere formulato nel seguente modo:

1. si suppone che θ_1 e θ_2 siano il vero effetto del trattamento sperimentale e di quello di controllo, rispettivamente;
2. e che $\delta = \theta_1 - \theta_2$ sia la vera differenza dei due effetti.

Allora il sistema di ipotesi da verificare è:

$$\begin{cases} \mathrm{H}_0: \theta_1 \leq \theta_2 - \delta_0 \\ \mathrm{H}_1: \theta_1 > \theta_2 - \delta_0, \end{cases}$$

ossia l'ipotesi alternativa H_1 è che il nuovo farmaco è non inferiore a quello di controllo di una quantità precisa pari a δ_0, e quindi corrisponde all'ipotesi di non-inferiorità del farmaco sperimentale rispetto al farmaco di controllo. Quindi δ_0 sarà il margine di non inferiorità.
Il test è basato sulla seguente statistica test

$$Z^* = [\hat{\delta} - (-\delta_0)]/\widehat{SE}_{\hat{\delta}},$$

dove $\widehat{SE}_{\hat{\delta}}$ è il valore stimato dello standard error di $\hat{\delta}$. La regola per rifiutare l'ipotesi di inferiorità H_0 è: $z^*_{oss} > z_{1-\alpha}$, con $\alpha = 0.05$ dato che il test è unilaterale.
Questo implica che

$$\hat{\delta} - z_{1-\alpha} * \widehat{SE}_{\hat{\delta}} > -\delta_0,$$

e quindi si può utilizzare anche l'intervallo di confidenza per verificare la non-inferiorità: si rifiuta H_0 se il limite di confidenza inferiore (LCL) della vera differenza $\delta = \theta_1 - \theta_2$ supera $-\delta_0$.

Per riassumere ci sono dei vantaggi nell'usare il metodo basato sui ranghi. In primo luogo, la stima di $\hat{\theta}$ ottenuta minimizzando la funzione combinata di dispersione $D(\theta)$ sarà più robusta rispetto alla stima calcolata con metodi parametrici dato che l'influenza dei valori anomali è minore. In secondo luogo, non è richiesta l'assunzione di normalità degli errori e delle componenti della varianza e, infatti, le componenti della varianza non sono coinvolte nella stima di θ. Infine, permette di risolvere problemi che, attraverso metodi parametrici, sarebbero stati abbandonati.

2.4 Test in R

In questo paragrafo vengono presentate delle funzioni dei pacchetti "EQUIVNONINF" e "PowerTOST" usati con il software R per calcolare ciò che serve per eseguire test di non-inferiorità [2].

Il primo test in R presentato del pacchetto EQUIVNONINF è la funzione che calcola la potenza con il test esatto di Fisher per test di non-inferiorità. In particolare, la funzione calcola i valori esatti della potenza di due distribuzioni binomiali:

```
bi2ste1(m, n, eps, alpha, p1, p2),
```

che prende in input le dimensioni dei due campioni (*m, n*), il margine di non inferiorità rispetto all'odds ratio (eps), il livello di significatività (alpha) e le probabilità di successo per i due campioni (p1, p2). Il test d'ipotesi sarà:

$$\begin{cases} H_0: p1 - p2 \leq eps \\ H_1: p1 - p2 > eps. \end{cases}$$

Per esempio, implementando la funzione, nel software R, nel seguente modo il risultato risulta essere:

```
>bi2ste1(m=106, n=107, eps=0.5, alpha=0.05, p1=0.9245,
p2=0.9065)

# M = 106     N = 107     EPS = 0.5

# ALPHA = 0.05     P1 = 0.9245     P2 = 0.9065

# POWNR = 0.5055587          POW = 0.5799595
```

Quindi l'ouput dice che il test ha una potenza del test randomizzato pari a 0.505, mentre pari a 0.58 se il test è non-randomizzato.

Invece, se si ha bisogno di conoscere la dimensione dei due campioni, la funzione che può aiutare è quella che calcola la dimensione dei campioni calcolate con il test esatto di Fisher per test di non-inferiorità:

```
bi2ste2(eps, alpha, p1, p2, bet, qlambd),
```

che prende in input il margine di non inferiorità rispetto all'odds ratio `(eps)`, il livello di significatività `(alpha)` e le probabilità di successo per i due campioni `(p1, p2)`, il valore target della potenza `(bet)` e il rapporto tra le dimensioni *m/n* `(qlambd)`. Per esempio, implementando la funzione nel seguente modo il risultato risulta essere:

```
> bi2ste2(eps=0.5, alpha=0.05, p1=0.9245, p2=0.9065,
bet=0.80, qlambd=1.00)

 # EPS = 0.5    ALPHA = 0.05   P1 = 0.9245

 # P2 = 0.9065 BETA = 0.8    LAMBDA = 1   M = 194

 # N = 194          POW = 0.8010046
```

Quindi l'output fa vedere che la dimensione M minima per il primo campione è pari a 194 e lo stesso vale per il secondo campione ($N = 194$).

Con il package "`PowerTOST`", invece, per stimare la dimensione del campione esiste un'altra funzione:

```
sampleN.noninf(alpha, targetpower, logscale, margin,
        theta0, CV, robust, details, print),
```

che prende in input il livello di significatività `(alpha)`; il valore della potenza che deve essere compresa tra 0 e 1 `(targetpower)`; l'argomento `logscale` che viene specificato `TRUE` se si vuole che i dati vengano usati con una trasformazione logaritmica, `FALSE` altrimenti; l'argomento θ_0 che è un rapporto se `logscale = TRUE` (di default è pari a 0.95) o una differenza se `logscale = FALSE` (di default è pari a

-0.05); il margine di non inferiorità `(margin)` espresso in rapporto se `logscale = TRUE` (di default uguale a 0.8), mentre se `logscale = FALSE` è espresso come differenza di medie (di default uguale a -0.2); il coefficiente di variazione `(cv)`: dato come rapporto se `logscale = TRUE`, mentre se `logscale = FALSE` si riferisce alla deviazione standard residua della risposta; i gradi di libertà `(robust)` di default `FALSE` (cioè che vengono utilizzati i gradi di libertà indicati), se viene impostato su `TRUE` verranno utilizzati i gradi di libertà in base allo stimatore di base di Senn. Infine, ci sono dei comandi di grafica come `details` e `print` che se impostati su `TRUE` stampano i risultati.
Le ipotesi per un test di non-inferiorità sono:

I. se `logscale = TRUE`

$$\begin{cases} \mathrm{H}_0: \theta_0 \leq log(margin) \\ \mathrm{H}_1: \theta_0 > \log(margin) \end{cases}$$

II. se `logscale = FALSE`

$$\begin{cases} \mathrm{H}_0: \theta_0 \leq margin \\ \mathrm{H}_1: \theta_0 > margin. \end{cases}$$

Di seguito un esempio.
-Si vuole stimare la dimensione del campione con un coefficiente di variazione pari a 0.3:

```
> sampleN.noninf(CV = 3)
++++++++++++ Non-inferiority test ++++++++++++
            Sample size estimation
---------------------------------------------
Study design: 2x2 crossover
log-transformed data (multiplicative model)
alpha = 0.025, target power = 0.8
Non-inf. margin = 0.8
```

```
True ratio = 0.95, CV = 0.3

Sample size (total)
 n     power
  48     0.801658
```

-Si vuole stimare la dimensione del campione con un coefficiente di variazione pari a 0.3, però specificando margine e θ_0 come reciproci rispetto ai valori del primo esempio:

```
> sampleN.noninf(CV = 0.25, margin = 1.25, theta0 = 1/0.95)
++++++++++++ Non-superiority test ++++++++++++
Sample size estimation
---------------------------------------------
Study design: 2x2 crossover
log-transformed data (multiplicative model)
alpha = 0.025, target power = 0.8
Non-inf. margin = 1.2

True ratio = 1.052632, CV = 0.25

Sample size (total)

n     power

36   0.820330
```

Il risultato è lo stesso, $n = 36$ e la potenza del test pari a 0.820330.
Se si prova ad usare un altro tipo di approccio in R (*Bracketing Approach*) si noterà che i risultati saranno diversi se prima si usa il margine di non-inferiorità pari a 0.83 e il valore di θ_0 pari a 0.95 e poi i loro reciproci.

Capitolo 3: Conclusioni

Per applicare ad uno studio clinico un test di non-inferiorità, ossia quando si vuole confrontare un farmaco o un trattamento che potrebbe essere messo in commercio ad un trattamento o farmaco già preesistente bisogna tener presente di alcuni concetti:

I. per prima cosa lo studio clinico non deve essere testato solo statisticamente, ma anche in termini sanitari e quindi il farmaco/trattamento proposto deve garantire lo stato di salute del paziente dato che un test di non-inferiorità potrebbe essere non etico;
II. considerare il limite di non-inferiorità $\delta\ (>0)$ il punto dello studio più importante in quanto sarà la soglia che ci premetterà di decidere se il farmaco è non-inferiore a quello standard oppure no [10];
III. infine, i metodi per risolvere il test e quindi decidere l'indirizzamento verso H_0 o H_1: sono stati presentati statistiche test di Blackwelder [13,14] e intervalli di confidenza [12]. Si è visto che è più efficiente utilizzare una statistica test per individuare la regione di accettazione e di rifiuto dell'ipotesi di non-inferiorità in presenza di costanza nel tempo dell'effetto del farmaco di controllo rispetto agli intervalli di confidenza, se però, mancano i presupposti gli intervalli di confidenza risultano essere metodi più robusti rispetto a questo aspetto. Inoltre, sono stati presentati due test non parametrici e si è visto che il metodo TACT [15,16] risulta essere il preferito in quanto sia un buon compromesso rispetto alla problematica descritta sopra.

Capitolo 4:
Bibliografia e Sitografia

1. https://www.researchgate.net/
2. R: https://www.r-project.org/
3. G. Coppi, L. Cottini, G. Fedele, (2015), *Studi di non-inferiorità: considerazioni su impostazione e utilità*, Farmacia industriale;
4. EMA (*European Medicines Agency*): Agenzia europea per i medicinali protegge e promuove la salute dei cittadini e degli animali valutando e monitorando i medicinali all'interno dell'Unione europea (UE) e dello Spazio economico europeo (SEE);
5. AIFA: Agenzia Italiana del Farmaco;
6. Robert Temple, MD, Center for Drug Evaluation and Research, U.S. Food and Drug Administration, (2000), *Placebo-Controlled Trials and Active-Control Trials in the Evaluation of New Treatments. Part 1: Ethical and Scientific Issues*;
7. Fisher, L.D., Gent, M., Bhller, H.R. (2001), *Active-control trials: How would a new agent compare with placebo? A method illustrated with clopidogrel, aspirin, and placebo*, American Heart Journal, 141, 26-32;
8. Hasselblad V. and Kong, D.F. (2001). *Statistical methods for comparison to placebo in active-control trials*, Drug Information Journal, 35: 435-449;
9. FDA: La *Food and Drug Administration* è la più antica agenzia governativa americana per la protezione dei consumatori, utilizzando l'analisi chimica per monitorare la sicurezza dei prodotti agricoli. La FDA è un'agenzia all'interno del *Department of Health and Human Services*. https://www.fda.gov;

10. Turki A., Althunian, Anthonius de Boer, Rolf H. H. Groenwold and Olaf H. Klungel, *Defining the noninferiority margin and analysing noninferiority: An overview*, Br J Clin Pharmacol, 83: 1636–1642;
11. Snapinn S., Jiang Q., *Preservation of effect and the regulatory approval of new treatments on the basis of non-inferiority trials*. Stat Med 2008; 27: 382–91;
12. H. M. James Hung, Sue-Jane Wang, Yi Tsong, John Lawrence, Robert T. O'Neill, USA; *SOME FUNDAMENTAL ISSUES WITH NON-INFERIORITY TESTING IN ACTIVE CONTROLLED TRIALS*;
13. Francesca Solmi, *Il problema della non-inferiorità in statistica medica*, anno accademico 2006/2007, tesi di laurea triennale;
14. Larry Laster, Mary Johnson, *Non inferiority trials: the ' at least as good as ' criterion,* January 2003, Statistics in Medicine , 22:187-200;
15. R. Arboretti Giancristofaro, S. Bonnini, F. Solmi, *Non-parametric two-stage active control testing method for non-inferiority tests*, Quaderni di Statistica Vol. 10, 2008.
16. S.J. Wang, H.M. James Hung, *TACT method for non-inferiority testing in active controlled trials*, December 2002, Statistics in Medicine, 22:227-238;
17. M. Mushfiqur Rashid, 2003*, Rank-based test for non-inferiority and equivalence hypothesis in multi-centre clinical trial using mixed models*, 22:291-311;
18. European Medicines Agency, Committee for Medicinal Products for Human Use. *Guideline on the pharmacokinetic and clinical evaluation of modified release dosage forms.* London. 20 November 2014.