UNIVERSITA' DEGLI STUDI DI PADOVA

FACOLTA' DI SCIENZE STATISTICHE

TESI DI LAUREA TRIENNALE

IL PROBLEMA DELLA NON-INFERIORITA' IN STATISTICA MEDICA

Relatore: Pesarin Fortunato

Laureanda: Solmi Francesca

ANNO ACCADEMICO 2006/2007

INDICE

INTRODUZIONE

In sede di esperimenti clinici il problema di confronto tra popolazioni è sicuramente uno dei più affrontati; in particolare la questione risulta cruciale quando si ha a che fare con il confronto tra farmaci di nuova produzione e farmaci già esistenti, per la commercializzazione dei primi.

In quest'ambito di ricerca sta prendendo piede la questione riguardante la scelta della direzionalità dei test da utilizzare: a seconda delle diverse situazioni e realtà cliniche diviene necessario non sottovalutare questo tipo di decisione preliminare, cruciale per una corretta gestione dello studio che si sta conducendo; esiste infatti, come vedremo, la possibilità di applicare, oltre ai test non-direzionali che spesso però non sono di grande utilità per gli obiettivi che solitamente ci si pone all'inizio di questo tipo di studi, o i test direzionali ad una coda, come i test di non-inferiorità, o i test direzionali a due code, questi ultimi da molti in diverse circostanze, come vedremo, preferiti. Questa diversità-dualità esistente tra gli ultimi due test è principalmente legata all'errore del terzo tipo, tenuto in considerazione solamente dai test a due code, con però perdita di potenza.

In questo lavoro verrà allora dedicato un primo capitolo alla trattazione dei test direzionali ad una coda, ed in particolare ai test di non-inferiorità, all'interno del quale verranno affrontate diverse questioni legate a questo tipo di approccio. In seguito un secondo capitolo sarà dedicato ai test direzionali a due code, dove verrà introdotto il concetto di errore del III tipo e verranno discussi pregi e difetti di questo tipo di approccio, a seconda dei diversi possibili ambiti di applicazione, anche in un confronto con i test di non-inferiorità, di cui il test direzionale a due code costituisce una recente alternativa.

CAPITOLO 1

I TEST DI NON-INFERIORITA'

Nell'ambito del confronto tra diversi trattamenti medici sono diverse le questioni che devono essere affrontate: l'organo che in Europa si occupa di stabilire le regole di comportamento da tenere è l'EMEA[1].

L'obiettivo principale che ci si deve porre all'inizio di uno studio clinico è quello di stabilire se il nuovo farmaco che si intende commercializzare dia non solo risultati migliori (in termini di efficacia e sicurezza) del placebo, ma abbia anche prestazioni non inferiori a quelle date dal migliore farmaco in commercio al momento del test. I motivi di questi obblighi sono evidentemente etici e gli esperimenti clinici sull'uomo sono oggi protetti da severe normative di carattere etico-morale; tutto questo crea senza dubbio una serie di problematiche legate al tipo di test che rimane possibile applicare nelle diverse situazioni, dalla più semplice in cui si dispone di tre popolazioni e quindi dei dati relativi a placebo, farmaco di controllo e farmaco testato, alla più problematica, in cui non è eticamente possibile somministrare il placebo, oppure non si dispone di dati storici affidabili sul confronto tra questo e il trattamento di controllo.

E' proprio a questo punto che è necessario affrontare la questione sulla scelta del test da utilizzare: il concetto di non-inferiorità accennato poco fa è il punto cruciale, in quanto non si parla più allora di test per testare la superiorità di un trattamento rispetto ad un altro, il classico test direzionale ad una coda che sta alla base dell'inferenza statistica classica, ma di test di non-inferiorità in cui si considera come accettabile per l'effetto del trattamento sperimentale una soglia inferiore all'effetto del trattamento di controllo. Questo tipo di approccio diverso da quello tradizionale è dettato sicuramente dalla particolarità del suo ambito di applicazione: quando si ricerca una cura ottimale per una determinata patologia si procede spesso a piccoli passi, perché spesso si cerca di premiare anche il più piccolo guadagno in termini di sicurezza per esempio, o di costi di

[1] L'EMEA (European Agency for the Evaluation of Medicines for Human Use) viene creata dall'Unione Europea nel 1995: essa ha il compito di valutare e supervisionare tutti i farmaci per uso umano e veterinario, con l'obbiettivo di creare una serie di regole standard per la ricerca e il commercio in questo campo.

produzione, in cambio di una piccola perdita di efficienza.

L'approccio da seguire per l'applicazione di questo tipo di test è quello di specificare un margine di non-inferiorità (diverso per ogni caso) con cui confrontare, con metodi di cui parleremo in seguito, la differenza tra due popolazioni trattate con i due diversi metodi di cura. La scelta di questo margine allora rappresenta una questione di massima importanza, in quanto, come è facile comprendere, essa influisce fortemente sulla decisione statistica che si andrà a prendere; essa deve tenere conto, come vedremo, tanto della significatività statistica quanto della rilevanza clinica relativamente agli specifici casi.

Un'altra delicata questione è sicuramente quella legata alla scelta della metodologia statistica da utilizzare, in quanto può essere più conveniente l'uso di intervalli di confidenza piuttosto che di test per verifiche di ipotesi in alcuni casi, come può diventare in altri (molti) casi necessario l'utilizzo di test non parametrici nel caso di impossibilità di assumere distribuzioni note per i dati a disposizione.

Le questioni accennate fin qui verranno affrontate nei vari paragrafi in modo separato, anche se naturalmente inevitabili saranno collegamenti e riferimenti tra le diverse sezioni del lavoro.

Infine verrà presentato uno studio di simulazione in cui si mostrerà la non distorsione del test in oggetto, attraverso un grafico e la relativa tabella di valori.

1.1. LA STRUTTURA DELL' ESPERIMENTO CLINICO

1.1.1. Condizioni generali

Come accennato sopra, l'EMEA dà precise disposizioni sul modo in cui l'esperimento clinico ideale per il confronto tra popolazioni trattate con farmaci diversi, allo scopo della commercializzazione del farmaco testato, dovrebbe essere strutturato; innanzitutto vi sono due assunzioni di base che devono essere necessariamente rispettate:

- l'effetto del trattamento deve essere misurabile;
- deve essere possibile distinguere tra un effetto positivo e uno negativo.

Risulta chiaro dagli scritti dell'EMEA, e appare comunque logico, che condizione necessaria perché un qualunque prodotto medico-farmaceutico possa essere approvato per la commercializzazione è che esso abbia un qualche effetto nel trattamento per cui è stato formulato in termini di rapporto rischi-benefici. Appare chiaro come sia necessario che il prodotto testato apporti al paziente alcuni benefici, in assenza dei quali invece non è possibile una decisione positiva in merito all'approvazione; da queste considerazioni segue che per dimostrare la propria efficacia il prodotto deve risultare significativamente superiore il termini di benefici al placebo.

Questo tipo di analisi viene solitamente condotta attraverso un test bilaterale ad un livello di significatività pari a 0.05 o un test unilaterale ad un livello pari a 0.025, oppure analogamente attraverso un intervallo di confidenza a due code al 95% o ad una coda al 97.5%.

Inoltre 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 questo, cosa che deve essere effettuata attraverso un test di non inferiorità o di superiorità a seconda dei singoli casi: la scelta di quale tra i due test utilizzare deve naturalmente essere pensata in modo attento congiuntamente all'analisi del contesto clinico in cui si sta operando.

In questo capitolo verranno descritte le due situazioni in cui ci si può trovare quando si intraprende uno studio clinico per testare l'efficacia di una trattamento

medico: il caso in cui si hanno a disposizione soggetti trattabili con placebo, con farmaco di controllo e con prodotto sperimentale, caso preferito ma non facilmente riscontrabile nella realtà, e un secondo caso in cui non si ha la possibilità di avere dati diretti riguardanti il placebo, e diviene allora necessario un confronto indiretto con essi attraverso una ricerca nella letteratura, caso per il quale si descriverà la procedura da seguire e le accortezze da avere durante tutto l'iter dello studio.

Verrà inoltre descritta l'esistenza di casi limite, come la non affidabilità dei dati reperiti attraverso la ricerca in letteratura o la totale assenza di dati relativi al placebo, e verrà quindi indicato il comportamento da tenere.

1.1.2. Placebo-controlled trials

La situazione ideale in cui ci si dovrebbe trovare è allora un esperimento in cui si hanno a disposizione tre popolazioni sulle quali vengono rilevate le variabili di interesse, una prima trattata con il placebo, una seconda con il farmaco di controllo, e una terza con il prodotto che si sta analizzando; avere a disposizione i dati relativi al placebo rende più efficienti gli esperimenti, richiedendo un minore numero minimo di pazienti per determinare l'effetto del trattamento da analizzare; qualora possibile, perché non impedito da questioni etiche o di altro genere, l'uso del trattamento placebo dovrebbe sempre essere adottato negli esperimenti clinici. In questo caso la procedura standard da seguire consiste nel testare la significatività statistica del trattamento di controllo rispetto al placebo. Una volta ottenuto un risultato positivo in questo senso si deve procedere per stabilire se la differenza registrata sia anche clinicamente rilevante: si deve condurre insomma un'analisi meno statistica e più clinica, riguardante anche la sicurezza del farmaco di controllo stesso e il suo rapporto rischi-benefici. A questo punto, una volta confermata anche la rilevanza clinica, si procede con il confronto tra il trattamento di controllo e il nuovo prodotto.

Particolare attenzione deve essere data al caso in cui il farmaco di controllo non fa registrare differenze significative rispetto al placebo, oppure si comporta in modo differente da quello che storicamente sarebbe naturale aspettarsi: in questo caso non si deve frettolosamente concludere che non esiste differenza

significativa tra controllo e placebo, ma si devono ipotizzare le possibili cause dell'anomalia, e trattare di conseguenza i successivi risultati ottenuti dal confronto tra controllo e nuovo trattamento.

1.1.3. Active-controlled trials

Come si può facilmente comprendere, la situazione in cui si hanno a disposizione i dati relativi ai tre trattamenti di interesse non è sempre semplice da trovare.

Può spesso accadere di non avere a disposizione il gruppo di pazienti trattato con il placebo: in questo caso diventa necessario ricorrere ad un confronto indiretto con altri dati relativi al placebo stesso. Deve essere effettuata un'attenta ricerca, nella letteratura a disposizione, dei dati relativi agli studi fatti sul confronto tra il placebo e il farmaco di controllo, al fine di arrivare ad una stima affidabile della vera differenza tra il farmaco e il placebo. In questa fase dello studio è necessario assicurarsi che siano soddisfatte alcune condizioni necessarie perché la ricerca nella letteratura e i risultati raggiunti con essa possano essere considerati affidabili:

- la scelta di quali tra gli studi riportati dalla letteratura considerare ai fini della nostra ricerca deve essere imparziale, e i criteri utilizzati nella scelta devono essere esplicitati in modo chiaro;
- deve esserci costanza nelle pratiche cliniche e nelle varie metodologie utilizzate nel tempo; devono allora essere considerati irrilevanti, e quindi non devono essere considerati nella ricerca, tutti quegli studi che sono stati condotti su campioni in cui le metodologie di somministrazione e conduzione dell'esperimento, e di raccolta dei dati e misurazione delle performance dei trattamenti non erano le stesse utilizzate all'interno dell'esperimento che si sta conducendo;
- le performance del farmaco di controllo devono essere costanti nel tempo: se si rilevano cambiamenti tra i vari studi è necessario considerare nella ricerca solamente gli studi più recenti, e se nemmeno con questi ultimi si riesce a garantire la costanza nel tempo, è necessario tenerne conto

successivamente nella scelta del margine di non-inferiorità per il successivo confronto tra trattamento di controllo e nuovo prodotto;

- è bene fare attenzione all'imparzialità delle pubblicazioni presenti in letteratura: se si pensa che possa esserci stata una certa tendenza a pubblicare più risultati positivi in favore del farmaco di controllo è opportuno lavorare successivamente con una ragionevole sottostima dei risultati registrati.

A questo punto, rispettate a dovere le accortezze sopra riportate, se si riesce ad avere dai dati storici una stima soddisfacente in termini di affidabilità della vera differenza, statisticamente significativa, tra placebo e farmaco di controllo, di nuovo come nel caso più semplice citato sopra, la questione da affrontare diventa stabilire se questa differenza esistente sia anche clinicamente rilevante. Ancora, una volta stabilita la rilevanza clinica del farmaco di controllo si deve procedere con l'esperimento per analizzare le performance del nuovo farmaco messe a confronto con i dati relativi a placebo e farmaco di controllo: a questo punto appare logico come, assunta l'assenza di problemi di sicurezza per il nuovo prodotto, si debba considerare a sua volta clinicamente rilevante e statisticamente significativo il guadagno del nuovo farmaco in termini di prestazioni rispetto al placebo, nel caso in cui questo risulti statisticamente superiore al farmaco di controllo, senza la necessità di condurre un confronto diretto tra i due. Inoltre la conoscenza in termini di grandezza della differenza di performance tra placebo e farmaco di controllo permette anche di pesare nel modo corretto le differenze tra quest'ultimo e il nuovo trattamento, dando eventualmente rilevanza clinica ad effetti che, all'oscuro di queste utili informazioni, sarebbero stati considerati di nessun peso.

In questi esperimenti a questo punto dell'analisi è possibile applicare due tipi di test per testare l'efficacia del nuovo trattamento su quello di controllo:

- se lo sponsor ha ragione di credere che il nuovo prodotto sia superiore al controllo, allora diventa giustificato un semplice test di superiorità per dimostrare che il farmaco sperimentale è superiore statisticamente al controllo;
- se non è ipotizzabile la suddetta superiorità allora diventa più appropriato applicare un test di non-inferiorità; questo stesso tipo di test può essere saggiamente adottato anche nel caso in cui sia da escludere un

miglioramento di prestazioni del farmaco sperimentale rispetto al controllo, però si abbia un sostanziale sorpasso del primo sul secondo in termini di sicurezza.

Anche qui una particolare attenzione deve però essere data al caso in cui il trattamento di controllo, risultato statisticamente e clinicamente superiore al placebo, in sede di esperimento fornisca dati che si discostano particolarmente dai dati storici registrati e riportati dalla letteratura considerata durante l'analisi: in questa situazione le performance fornite dal controllo potrebbero non essere affidabili, e di conseguenza i risultati relativi al confronto tra controllo e nuovo trattamento non possono essere utilizzati per fare inferenza su tutta la popolazione statistica, ma devono essere considerati casi particolari, e possono essere discusse le eventuali possibili cause dei valori anomali.

E' infine necessario specificare che in alcuni campi, come l'oncologia o la pediatria, l'utilizzo del placebo per esperimenti clinici viene considerato assolutamente non etico: in questi casi, se non si hanno a disposizione dati utili alla costruzione di un intervallo di confidenza indiretto per la differenza tra nuovo trattamento e placebo, si deve considerare come placebo il trattamento di controllo, e allora andare a condurre necessariamente un test di superiorità del nuovo sul controllo.

1.2 LA SCELTA DEL MARGINE DI NON-INFERIORITA'

1.2.1. La scelta del margine di non-inferiorità: una questione delicata

Come abbiamo anticipato, la scelta del margine di non-inferiorità è un'operazione molto delicata, che dovrebbe essere condotta utilizzando congiuntamente risultati di studi statistici e conoscenze cliniche del caso in questione, e cercando di mantenere un comportamento corretto dal punto di vista etico e morale.

Di seguito verranno dapprima analizzati il problema e la struttura delle ipotesi da testare, in modo da definire formalmente il margine in questione e capire così il ruolo che esso copre e la delicatezza della sua posizione dal punto di vista non solo statistico, ma anche e soprattutto etico e morale legato alle possibili conseguenze mediche sui soggetti trattati nello studio.

Infine verranno presentate la procedura e le regole da seguire per una corretta scelta del margine stesso.

1.2.2. La formulazione delle ipotesi

Definiti T il valore della variabile di interesse (indicatrice dell'efficacia) misurata per il trattamento sperimentale e C la misura relativa al trattamento di controllo, e supponendo che siano desiderabili valori alti della suddetta variabile, allora le ipotesi standard del test di non-inferiorità sono quelle descritte di seguito:

$$\begin{cases} H_0: \quad C - T \geq \delta & \text{(C è superiore a T)} \\ H_1: \quad C - T < \delta & \text{(T è non inferiore a C)} \end{cases}$$

dove δ (>0)indica il margine di non-inferiorità[2].

L'interpretazione di δ allora appare molto semplice: esso descrive di quanto al massimo T può essere inferiore a C per continuare ad essere considerato non inferiore. L'ipotesi nulla se non rifiutata porta ad affermare la superiorità di C rispetto a T, mentre nel caso contrario si è autorizzati ad affermare la non

[2] D'Agostino R. B.; "Non-inferiority trias: design concepts and issues – the encounters of academic consultants in statistics", in *Statistics in Medicine*, (pp 173-174); 2003.

inferiorità di T rispetto a C, formalizzata quindi dall'ipotesi alternativa H_1.

Il problema della scelta tra le due ipotesi può essere risolto con un test unilaterale ad un livello di significatività α, oppure tramite un intervallo di confidenza simmetrico al un livello del (1-2α) per la differenza ($C - T$), confrontando il limite superiore dell'intervallo con il valore del margine δ, rifiutando H_0 laddove questo limite risulti minore di δ[3].

Risulta allora chiaro come la scelta del margine δ sia una questione di primaria importanza per arrivare ad un'affidabile decisione in merito alla verifica di ipotesi sopra descritta.

1.2.3. La scelta di δ

L'EMEA avverte che la scelta del margine di non-inferiorità δ deve essere fatta nel modo più prudente e attento possibile: i test di non-inferiorità vengono condotti all'interno degli esperimenti clinici al fine di individuare e valorizzare alcuni farmaci che con un test di superiorità verrebbero bocciati, e che invece potrebbero essere tranquillamente commercializzati in quanto peggiori in termini di efficienza di una quantità trascurabile, ma spesso migliori in termini di sicurezza. Insomma si vuole in questo modo ammettere un'eventuale leggera perdita di efficacia a favore di un guadagno in sicurezza, e in merito a questo è giustificato un aumento del margine δ in casi in cui ci sia un miglioramento consistente in termini di sicurezza.

Nel caso particolare di mercati in cui è presente un solo prodotto in commercio, la scelta di δ dovrà essere condotta anche tramite l'ausilio delle risposte di pazienti in merito alla massima perdita di efficienza da essi considerata ammissibile in quanto non importante, naturalmente con l'accorgimento di controllare che le risposte non siano influenzate verso valori grandi di δ.

In generale per tutti i mercati di interesse la scelta di δ deve essere basata su obiettivi precisi, che devono essere decisi tenendo presenti i dati storici relativi al confronto tra trattamento di controllo e placebo ottenuti dalla ricerca nella letteratura e singolarmente per ogni studio clinico.

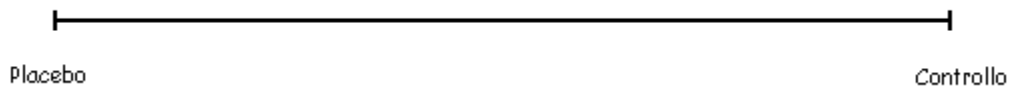
[3] D'Agostino R. B.; "Non-inferiority trias: design concepts and issues – the encounters of academic consultants in statistics", in *Statistics in Medicine*, (p 174); 2003.

Le regole principali da tenere in considerazione per la determinazione di δ sono essenzialmente due[4]:

- la scelta deve essere basata su ragionamenti sia statistici che clinici, al fine di dare il giusto peso alla parte prettamente medica del problema, che non deve essere assolutamente trascurata; in realtà molto di rado nei casi pratici sembra venire tenuto conto della necessità di questi aggiustamenti clinici nella determinazione del margine, che sembra piuttosto essere fatta, come già accennato, quasi esclusivamente secondo studi statistici basati sui dati storici a disposizione dalla letteratura. E' chiaro allora come di conseguenza il margine di non inferiorità scelto facendo riferimento ai dati storici rifletta inevitabilmente l'incertezza che caratterizza gli stessi;
- il margine scelto non deve essere superiore alla più piccola differenza di effetto registrato in sede di confronto tra farmaco di controllo e placebo, al fine di garantire la superiorità del trattamento sperimentato sul placebo stesso.

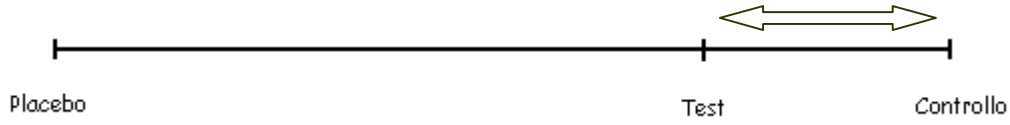
Il processo di determinazione della non-inferiorità all'interno di uno studio clinico può essere comunque riassunto nei seguenti passi fondamentali[5]:

a) innanzitutto è necessario assicurarsi della superiorità statistica del farmaco di controllo sul placebo, attraverso i dati storici disponibili; una volta dimostrata la superiorità bisogna però fare una forte assunzione di costanza di questa nel tempo, che permette di affermare che essa persiste anche nello studio in oggetto; questo corrisponde ad assumere una certa efficacia del farmaco di controllo rispetto al placebo nel presente studio:

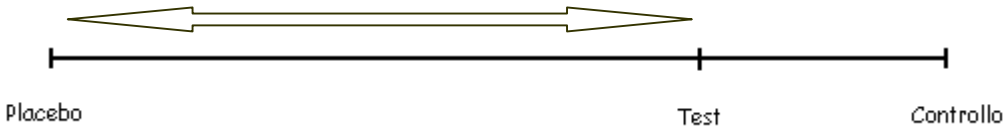


[4] D'Agostino R. B.; "Non-inferiority trias: design concepts and issues – the encounters of academic consultants in statistics", in *Statistics in Medicine*, (p 174); 2003.
[5] D'Agostino R. B.; "Non-inferiority trias: design concepts and issues – the encounters of academic consultants in statistics", in *Statistics in Medicine*, (pp 174-175-176); 2003.

b) si deve poi dimostrare che il trattamento sperimentale ha un'efficacia che rientra all'interno di un certo margine, che naturalmente deve avere una rilevanza clinica:



c) infine si devono utilizzare congiuntamente i dati relativi al confronto tra T e C e tra C e P (dove P indica il valore della variabile di interesse misurata per il placebo) per dimostrare indirettamente la superiorità di T su P; inoltre è necessario dimostrare che non solo T è superiore a P, ma che esso mantiene una certa percentuale, che in seguito indicheremo con X, sufficientemente alta e decisa a seconda dei casi tra il 50 e l'80%, dell'effetto del farmaco di controllo sul placebo:



A questo punto, com'è facile intuire, il margine δ deve essere allora fissato di conseguenza alla decisione della percentuale di efficacia che si vuole sia mantenuta dal trattamento sperimentale: il margine di non inferiorità deve essere fissato non superiore alla quantità (1-X)(C-P).

Vi possono essere diversi modi per determinare il margine di non-inferiorità, ma ad oggi nessuno di essi è preferito agli altri. Di seguito verranno descritte alcune procedure basate sul concetto di rischio relativo[6]: chiamati T, C e P i rapporti di incidenza di un evento clinico nella popolazione da cui sono stati campionati i pazienti trattati rispettivamente con il farmaco sperimentale, quello di controllo e con il placebo, e C_0 e P_0 quelli relativi rispettivamente al controllo

[6] Hung H. M. J., Wang S., Tsong Y., Lawrence J., O'Neil R. T.; "Some fundamental issues with non-inferiority testing in active controlled trials", in *Statistics in Medicine*, (pp 215-216); 2003.

e al placebo nei dati storici trovati in letteratura, se l'effetto del trattamento viene espresso in termini di rischio relativo, l'ipotesi di non-inferiorità diviene:

$$\begin{cases} H_0: & T/C \geq \delta \\ H_1: & T/C < \delta \end{cases}$$

Allora si può procedere nei seguenti modi:

- si considerano i due valori ottimali per il margine trovati tramite la ricerca in letteratura, e quindi da un punto di vista statistico, e tramite gli studi clinici, e si sceglie il più piccolo tra i due, in un'ottica più conservativa;
- si considera una certa percentuale 100λ dell'effetto del trattamento di controllo da conservare in termini di rischio relativo, e cioè:

  $$1 - T/P > \lambda(1 - C/P)$$

  o equivalentemente:

  $$T/C < 1 + (1-\lambda)(P/C - 1)$$

  dove λ, P, C, T sono compresi tra 0 e 1.

  Allora il margine di non-inferiorità è dato da:

  $$1 + (1-\lambda)(P/C - 1);$$

- si può ripetere il ragionamento fatto al punto precedente considerando il rischio relativo in termini di logaritmo, in quanto, come noto, questa trasformazione aiuta l'assunzione di normalità per le statistiche in oggetto; allora le due ipotesi diventano:

  $$\begin{cases} H_0^{|}: & \log(T) - \log(C) \geq \delta \\ H_1^{|}: & \log(T) - \log(C) < \delta \end{cases}$$

  o equivalentemente:

  $$\begin{cases} H_0^{|}: & [\log(P) - \log(T)]/[\log(P) - \log(C)] \leq \gamma \\ H_1^{|}: & [\log(P) - \log(T)]/[\log(P) - \log(C)] > \gamma \end{cases}$$

  dove allora il margine di non-inferiorità è dato da:

  $$\delta = (1-\gamma)[\log(P) - \log(C)]$$

Qualunque metodo si segua per la determinazione del margine di non-inferiorità è necessario tenere conto dell'incertezza che caratterizza la stima dell'effetto $(C - P)$, e spesso è opportuno optare per una scelta più conservativa, prendendo come stima il più basso dei limiti inferiori di tutti gli intervalli di confidenza per la quantità in questione forniti dai diversi studi disponibili in

letteratura; questo argomento verrà comunque discusso nei paragrafi successivi, insieme a tutte le questioni e le problematiche ad esso connesse.

E' infine necessario fornire alcuni avvertimenti relativamente al problema della scelta di δ[7], e quindi di conseguenza ad alcune questioni ad esso legate:

- vi sono alcuni campi medici, come quello riguardante gli anestetici, in cui non è sufficiente basarsi sui dati storici per stimare correttamente l'efficacia del farmaco di controllo sul placebo, in quanto esperimenti diversi possono dare risultati anche molto diversi tra di loro. In questi casi è necessario che nello studio che si sta conducendo sia presente un campione di pazienti trattato con placebo;
- è molto importante che venga sempre tenuta in considerazione l'assunzione di costanza nel tempo delle metodologie e degli strumenti utilizzati all'interno degli esperimenti, e che ne venga rigorosamente controllata la validità, soprattutto ai giorni nostri, caratterizzati da un continuo innovamento tecnologico in campo medico;
- la stima della quantità $(C-P)$ deve essere fatta in modo conservativo, come già accennato, per esempio prendendo il più piccolo tra i limiti inferiori dei diversi intervalli di confidenza disponibili, ma sempre facendo attenzione a prendere un valore che risulti statisticamente significativo, e che si riferisca a studi trovati in letteratura che rispettino tutte le condizioni citate al par. 1.1.3. relativamente ai dati storici affidabili;
- si possono trovare situazioni in cui vi è a disposizione in letteratura solamente uno studio che fornisce una stima dell'efficacia del farmaco di controllo sul placebo: in questi casi bisogna procedere con prudenza nelle varie fasi dello studio che si sta conducendo e tenerne conto in sede di interpretazione dei risultati, in quanto la stima ottenuta per la quantità $(C-P)$, da cui sappiamo essere fortemente influenzato l'intero esperimento, potrebbe risultare poco affidabile;
- situazione estrema ma presente in diversi campi di studio è quella in cui ci si trova in assenza totale di dati storici sul confronto tra trattamento di controllo e placebo; in questo caso si può (o meglio si è costretti a) considerare alla

[7] D'Agostino R. B.; "Non-inferiority trias: design concepts and issues – the encounters of academic consultants in statistics", in *Statistics in Medicine*, (pp 176-177); 2003.

stregua del placebo il trattamento di controllo somministrato a dosi basse, con tutti i rischi che, come si può facilmente intuire, ne possono derivare.

1.3 I TEST

1.3.1. Diversi approcci

Per risolvere il problema della non inferiorità di un nuovo prodotto sperimentale rispetto ad un certo farmaco di controllo, sono presenti in letteratura diverse proposte: qui ne riporteremo alcune, che differiscono molto tra di loro per condizioni di applicazione e metodologie utilizzate.

In generale il problema della non-inferiorità viene spesso in letteratura affrontato in termini di “at least as good as” criterion, ad intendere una procedura statistica basata su test o intervalli di confidenza strutturata per dimostrare la non-inferiorità di una terapia sperimentale rispetto al trattamento di controllo. Come abbiamo già specificato, non si tratta allora di dimostrare la superiorità, che peraltro è una condizione che non è esclusa nel caso di dimostrazione della non-inferiorità, ma si vuole semplicemente provare che il farmaco testato dà prestazioni di cura che sono almeno tanto efficaci quanto quelle del farmaco di controllo meno una quantità, il margine di non inferiorità appunto, che deve essere deciso secondo dei ragionamenti sia clinici che statistici, come abbiamo già visto nel corso del paragrafo precedente. Per applicare questo concetto, si trovano in letteratura dei test che mettono a confronto un’ipotesi nulla di superiorità della terapia di controllo rispetto al trattamento sperimentale, con l’ipotesi alternativa di non- inferiorità descritta sopra, test che sono già stati formalizzati nel capitolo precedente.

Un primo approccio che verrà approfondito si riferisce al test proposto da Blackwelder, che verrà affrontato sia in termini di differenza che di rapporto tra medie di popolazioni: si dimostrerà che questo secondo approccio basato sul rapporto è più efficiente di quello proposto da Blackwelder, in quanto esso richiede una numerosità campionaria inferiore rispetto a quest’ultimo per individuare con una certa potenza la soluzione del problema.

Un secondo approccio che verrà analizzato è il cosiddetto metodo TACT (Two-stage Active Control Testing): come sarà mostrato, i due possibili metodi standard per testare la non-inferiorità in studi clinici sono il metodo di sintesi, che utilizza una statistica test per individuare le regioni di accettazione e di

rifiuto dell'ipotesi nulla, e il metodo degli intervalli di confidenza; il primo risulta più efficiente sotto l'assunzione di costanza nel tempo dell'effetto del farmaco di controllo di cui si è accennato nel capitolo 2, ma quando vengono a mancare i presupposti per quest'assunzione vedremo che l'errore del primo tipo aumenta rapidamente, compromettendo l'affidabilità dell'intero studio. Il secondo approccio invece risulta più robusto rispetto a quest'aspetto. Il metodo TACT è stato sviluppato come valido compromesso tra i due diversi approcci. Alla definizione dei due metodi standard e al problema della variabilità dell'errore del primo tipo verrà dedicato un paragrafo a parte per chiarire bene i termini della questione; un successivo paragrafo sarà dedicato nello specifico al metodo TACT.

Ai test basati sui ranghi verrà dedicata l'ulteriore sezione, nella quale verranno discusse le motivazioni per l'utilizzo di questo tipo di approccio in sostituzione a quello parametrico e i vantaggi che questo può portare.

L'ultima sezione di questo terzo paragrafo sarà infine dedicata ai risultati di uno studio di simulazione condotto sul test di non-inferiorità.

1.3.2. Il test di Blackwelder

In ambito clinico quando si confrontano un trattamento sperimentale con uno standard di controllo si possono distinguere in generale quattro stati di natura (dal punto di vista del farmaco sperimentale sul controllo):

i. Inferiorità clinica;
ii. Tolleranza clinica;
iii. Equivalenza;
iv. Superiorità.

La frase "at least as good as", che indica uno stato di non-inferiorità appunto, può essere identificata con l'unione degli ultimi tre stati di natura; allora il test proposto da Blackwelder è stato strutturato proprio per studiare questo tipo di ipotesi.

In questa trattazione l'approccio fornito da Blackwelder verrà affrontato come già accennato anche in termini di rapporto tra medie. Prima di concentrarci su questa estensione, definiremo con precisione il test basato sulla differenza tra

medie, in modo da inquadrare e capire al meglio i due diversi tipi di approccio e poter poi fare gli adeguati confronti.
Il test proposto da Blackwelder studia le seguenti ipotesi:

$$\begin{cases} H_0: & \mu_C - \mu_T \geq \delta \\ H_1: & \mu_C - \mu_T < \delta \end{cases}$$

dove: μ_C = media della variabile indicatrice dell'efficacia del farmaco di controllo,

μ_T = media della variabile indicatrice dell'efficacia del farmaco sperimentale,

δ = margine di non inferiorità, definito >0[8]

Bisogna specificare che il test è così formalizzato per studi nei quali incrementi del valore delle due medie denotano un miglioramento in termini di efficacia del farmaco; in caso contrario le ipotesi devono essere riformulate di conseguenza.
Analogamente allo schema proposto in precedenza, anche qui l'ipotesi nulla rappresenta il caso di superiorità statistica del farmaco di controllo su quello sperimentale, corrispondente al primo stato di natura presentato, mentre l'ipotesi alternativa rappresenta lo stato che abbiamo chiamato "at least as good as", cioè appunto l'unione degli ultimi tre stati di natura.
Più precisamente, per meglio comprendere il rapporto esistente tra le formali ipotesi e i vari stati di natura possiamo definire le seguenti corrispondenze, riportate nella tabella 1:

[8] Laster L. L., Johnson M. F.; "Non-inferiority trials: the 'at least as good as' criterion"; in *Statistics in Medicine*, (p 189); 2003.

Inferiorità clinica	Tolleranza clinica	Equivalenza	Superiorità
$\mu_C - \mu_T \geq \delta$	$0 < \mu_C - \mu_T < \delta$	$\mu_C - \mu_T = 0$	$\mu_C - \mu_T < 0$
La terapia sperimentale è inferiore a quella di controllo di una quantità δ o più	La terapia sperimentale è inferiore a quella di controllo di una quantità minore di δ	Le due terapie sono equivalenti	La terapia sperimentale è superiore a quella di controllo

Tabella 1: corrispondenze concettuali tra ipotesi formali del test di Blackwelder basato sulla differenza tra le medie delle due popolazioni e le diverse condizioni da un punto di visto clinico, la parte inferiore della tabella descrive le ipotesi statistiche riportate nella seconda riga.

1.3.3. La versione basata sul rapporto

Il test di Blackwelder lavora con la differenza assoluta esistente tra i due trattamenti considerati; l'approccio che verrà proposto di seguito, basando le proprie ipotesi sul rapporto tra le medie dei due trattamenti, si serve invece della differenza relativa esistente tra i due farmaci, portando una serie di vantaggi che verranno esplicitati in questo paragrafo.

Formalmente questo test studia le seguente ipotesi

$$\begin{cases} H_0: & \mu_T / \mu_C \leq R \\ H_1: & \mu_T / \mu_C > R \end{cases}$$

dove: R = limite inferiore di percentuale di efficacia del farmaco di controllo, definito <1[9].

Anche in questo caso, come nel paragrafo precedente, bisogna specificare che il test è così formalizzato per studi nei quali incrementi del valore delle due medie denotano un miglioramento in termini di efficacia del farmaco; in caso contrario le ipotesi devono essere riformulate di conseguenza.

[9] Laster L. L., Johnson M. F.; "Non-inferiority trials: the 'at least as good as' criterion"; in *Statistics in Medicine*, (p 189-190); 2003.

Solitamente il limite R viene scelto alto (tra l'80 e il 90%) in modo da garantire che il trattamento sperimentale assicuri un effetto clinicamente rilevante. Come per il test di Blackwelder l'ipotesi nulla identifica lo stato di inferiorità clinica della terapia sperimentale rispetto a quella di controllo, mentre l'ipotesi alternativa è indicativa della non-inferiorità del farmaco sperimentale rispetto a quello standard.

I due test sono quindi analoghi dal punto di vista dell'interpretazione del rapporto esistente tra i due trattamenti, in termini di stati di natura (interpretazione delle ipotesi nulla e alternativa), però leggermente diversi nella specifica definizione del tipo di differenza esistente, come descritto in tabella 2:

Inferiorità clinica	Tolleranza clinica	Equivalenza	Superiorità
$\mu_T/\mu_C \leq R$	$R < \mu_T/\mu_C < 1$	$\mu_T/\mu_C = 1$	$\mu_T/\mu_C > 1$
La terapia sperimentale ha un'efficacia inferiore al $(R*100)\%$ dell'efficacia di quella di controllo	La terapia sperimentale ha una perdita di efficacia rispetto a quella di controllo al più del $((1-R)*100)\%$	Le due terapie sono equivalenti	La terapia sperimentale è superiore a quella di controllo

Tabella 2: corrispondenze concettuali tra ipotesi formali della versione del test di Blackwelder basata sul rapporto tra le medie delle due popolazioni e le diverse condizioni da un punto di visto clinico, la parte inferiore della tabella descrive le ipotesi statistiche riportate nella seconda riga.

Come possiamo vedere dalla tabella, con quest'approccio, oltre a cambiare il valore che identifica l'equivalenza tra i due trattamenti da 0 a 1 (ovviamente dato che si sta parlando di rapporto e non più di differenza), la cosa che più interessa è che i primi due stati di natura, e quindi le prime due regioni di valori per il test identificate dai punti critici R e 1, sono in grado di fornire una differenza relativa tra i due farmaci, quantità sicuramente di più facile interpretazione rispetto al suo corrispondente assoluto.

Quindi il vantaggio fondamentale offerto da questo tipo di approccio è proprio quello di vedere l'efficacia del trattamento sperimentale in termini di percentuale dell'efficacia del farmaco di controllo.

I due approcci vengono a coincidere nel caso in cui nel test di Blackwelder si scelga come valore per δ una piccola percentuale (allora per analogia compresa tra il 10 e il 20%) della media μ_C del trattamento di controllo., infatti si avrebbe in questo caso:

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

da cui seguirebbe che l'ipotesi nulla del test di Blackwelder potrebbe essere riscritta come:

$$\begin{aligned} H_0: \quad & \mu_C - \mu_T \geq \delta \\ & \mu_C - \delta \geq \mu_T \\ & \mu_C - (1 - R)\mu_C \geq \mu_T \\ & \mu_C - \mu_C + R\mu_C \geq \mu_T \\ & R\mu_C \geq \mu_T \\ & \mu_T / \mu_C \leq R \end{aligned}$$

essendo quest'ultima proprio l'ipotesi nulla del test basato sul rapporto tra le medie.

1.3.4. Metodi di decisione per il test sul rapporto tra le medie

Il test sopra descritto può essere affrontato tramite l'uso di una tradizionale statistica test unidirezionale da confrontare con il relativo valore critico ad un livello di significatività scelto a priori α, oppure attraverso la costruzione di un intervallo di confidenza simmetrico ad un livello (1 - 2α) per il vero valore del rapporto tra le due medie, che chiameremo R_{True} da confrontare con il valore prescelto per R.

Come stimatore per R_{True} deve essere considerato:

$$\tilde{R} = \overline{X}_T / \overline{X}_C$$

che è asintoticamente non distorto e normalmente distribuito, sotto le assunzioni:

- $X_{T(i)} \sim N(\mu_T, \sigma_T^2)$, $i = 1,...n$ *e* $X_{C(i)} \sim N(\mu_C, \sigma_C^2)$, *i = 1,…,n, variabili casuali indipendenti;*
- $\sigma_T^2 = \sigma_C^2 = \sigma^2$, cioè le due variabili sono omoschedastiche;
- $n_T = n_C = n$, cioè i due campioni hanno la stessa numerosità.

I due approcci utilizzabili sono allora:

- **<u>Statistica test standard</u>**: per quanto riguarda questo metodo ci si serve di una riparametrizzazione delle ipotesi nulla e alternativa come segue[10]:

$$\begin{cases} H_0: & \mu_T - R\mu_C \leq 0 \\ H_1: & \mu_T - R\mu_C > 0 \end{cases}$$

  e cioè considerando come decisore tra inferiorità e non inferiorità del trattamento sperimentale sul farmaco di controllo il segno della differenza esistente tra la media della variabile indicatrice dell'efficacia del prodotto sperimentale e la percentuale della media del prodotto di controllo che è deciso, attraverso la scelta di R, essere clinicamente necessaria da mantenere.

  In questo modo si riesce ad ottenere una distribuzione teorica più trattabile, e si procede con la costruzione del test non distorto uniformemente più potente, dato dalla seguente definizione:

$$T = \left(\overline{X}_T - R\overline{X}_C\right) \Big/ \left[s^2\left(1 + R^2\right)/n\right]^{1/2}$$

  dove s^2 è la varianza corretta stimata per i due campioni; questo stimatore ha una distribuzione teorica di una *t* di Student con (2n – 2) gradi di libertà.

- **<u>Intervallo di confidenza</u>**: il metodo basato sulla costruzione dell'intervallo di confidenza risulta più complicato del primo appena descritto; la costruzione dell'intervallo in questione viene basata sulla seguente quantità:

$$\left(\hat{R} - R\right) \Big/ \left[s^2\left(1 + R^2\right)/n\left(\overline{X}_C\right)^2\right]^{1/2}$$

  che non dovrebbe superare in valore assoluto la costante critica $t_{\alpha;2n-2}$. I valori di R che eguagliano il rapporto sopra riportato al valore critico appena citato sono le radici di una complessa equazione quadratica, ma

[10] Il riferimento bibliografico di seguito riportato è relativo a tutte le formulazioni presentate in questa sezione: Laster L. L., Johnson M. F.; "Non-inferiority trials: the 'at least as good as' criterion"; in *Statistics in Medicine*, (p 190-191-192); 2003.

per campioni di numerosità abbastanza elevata le radici sono approssimabili ai seguenti valori, che forniscono allora l'intervallo di confidenza approssimato per R_{True}:

$$\hat{R} \pm z_{\alpha} \left[s^2 \left(1 + \hat{R}^2\right) \Big/ n\left(\overline{X}_C\right)^2 \right]^{1/2}$$

In generale nella pratica viene usata più spesso questa approssimazione piuttosto che il metodo esatto.

La regola per rifiutare l'ipotesi nulla in favore di quella alternativa di non-inferiorità consiste nel verificare che l'intervallo di confidenza, esatto o approssimato, non solo non contenga il valore R, ma anche, molto importante, che esso giaccia interamente alla destra di R, e cioè contenga solo valori, per R_{True}, più grandi di R stesso.

In generale il metodo basato sul test standard è quello più efficiente, ed il motivo sta essenzialmente nella diversa stima che viene fatta per lo standard error dello stimatore $\hat{R}$: adottando il metodo dell'intervallo di confidenza infatti deve necessariamente essere che il valore di $\hat{R}$ sia più grande di R perchè l'intervallo che poi verrà costruito su di esso abbia qualche possibilità di escludere R e quindi rifiutare l'ipotesi nulla in favore di quella di non-inferiorità, e questo porta necessariamente ad una stima dello standard error per $\hat{R}$ maggiore di quella che si ottiene invece tramite il metodo classico del test, che viene fatta supponendo $\hat{R} = R$. Di conseguenza il valore del limite inferiore dell'intervallo di confidenza considerato potrebbe risultare poco affidabile.

Per il metodo standard del test forniamo di seguito allora la numerosità ottimale, per ciascun gruppo, ad un livello di significatività α e per una potenza del test posta pari a β:

$$n = \left[(CV)^2 \left(z_{1-\alpha} - z_{\beta}\right)^2 \left(1 + R^2\right) \right] \Big/ \left(R_{True} - R\right)^2$$

1.3.5. Confronto tra il test di Blackwelder e la sua versione rapporto

Mettendo a confronto il test di Blackwelder e la sua versione basata sul rapporto tra medie il secondo risulta apportare diversi benefici, sia dal punto di vista statistico, che da quello medico.

Per quanto riguarda l'aspetto statistico, che più ci interessa, nel caso in cui l'ipotesi alternativa H_1: $\mu_C - \mu_T < \delta$ o equivalentemente H_1: $\mu_T / \mu_C > R$ sia vera, e $\delta > 0$ o equivalentemente $R < 1$, si dimostra che il secondo approccio risulta avere un'efficienza maggiore, cioè richiede una numerosità dei campioni minore rispetto a quella richiesta dal test originale, a parità degli errori di primo e secondo tipo, fissati a priori.

Questo è essenzialmente dovuto alla differenza esistente tra gli errori standard che derivano dai due diversi approcci; infatti per quanto riguarda il test di Blackwelder esso si basa essenzialmente sulla seguente quantità[11]:

$$\overline{X}_C - \overline{X}_T - \delta$$

la cui varianza risulta pari a:

$$\mathrm{var}\left(\overline{X}_C - \overline{X}_T - \delta\right) = 2\sigma^2 / n$$

mentre il test basato sul rapporto tra le medie si basa, come abbiamo già descritto, sulla quantità:

$$\overline{X}_T - R\overline{X}_C$$

la cui varianza risulta quindi pari a:

$$\mathrm{var}\left(\overline{X}_T - R\overline{X}_C\right) = \sigma^2\left(1 + R^2\right) / n$$

da cui segue che, quando R < 1:

$$\mathrm{var}\left(\overline{X}_C - \overline{X}_T - \delta\right) > \mathrm{var}\left(\overline{X}_T - R\overline{X}_C\right)$$

Di conseguenza i due test, per R < 1, richiedono numerosità differenti come descritto.

Per quanto riguarda i vantaggi da un punto di vista puramente medico, un'ulteriore punto a favore del test basato sul rapporto è rappresentato dalla sua maggior semplicità di comprensione nonché di applicazione; esso è per

[11] Il riferimento bibliografico di seguito riportato è relativo a tutte le formulazioni presentate in questa sezione: Laster L. L., Johnson M. F.; "Non-inferiority trials: the 'at least as good as' criterion"; in *Statistics in Medicine*, (p 194); 2003.

questi motivi meglio accettato rispetto alla sua versione originale anche dal personale medico partecipante allo studio, che trova appunto più semplice definire una soglia di accettazione in termini di percentuale di una quantità già conosciuta, piuttosto che in termini di valore assoluto.
Inoltre la maggiore efficienza del test rapporto precedentemente dimostrata porta dei benefici che rientrano sicuramente anche nell'aspetto medico dello studio: essa significa maggiore potenza a parità di numerosità, e quindi minore numerosità richiesta a parità di potenza, il che significa minor numero di pazienti da includere nello studio, e quindi se vogliamo anche minori costi.

1.3.6. Due metodi a confronto: test standard contro intervalli di confidenza

Esistono nell'ambito delle soluzioni parametriche, come già descritto nei paragrafi precedenti, due possibili metodologie che possono venire utilizzate per risolvere il problema della non-inferiorità: la prima si basa sull'uso di una statistica test che viene confrontata con una distribuzione teorica e in particolare con un valore critico che ne decide l'appartenenza alla regione cosiddetta di accettazione o a quella di rifiuto dell'ipotesi nulla. La seconda metodologia si basa sulla costruzione di un intervallo di confidenza per il vero valore del parametro che si sta studiando (nel nostro caso la differenza, in termini assoluti o relativi, tra le performance di farmaco di controllo e di trattamento sperimentale) e sul confronto di questo con il valore deciso come soglia massima (o minima a seconda della tipologia del problema e delle ipotesi fatte).
Nei precedenti capitoli è stato fatto un confronto tra le due metodologie nel caso del test di Blackwelder basato sul rapporto tra medie in termini di efficienza dei due metodi, e si è visto come il primo metodo proposto risulti migliore del secondo.
Per un confronto completo tra i due metodi è necessario però affrontare un'altra questione che interessa direttamente l'entità dell'errore α del primo tipo e il suo rapporto con la presenza o meno della condizione di costanza di cui si è parlato nei capitoli precedenti, perché questo ha un impatto importante sull'affidabilità

delle conclusioni a cui si giunge tramite l'inferenza sui risultati ottenuti dallo studio. Il problema sta essenzialmente nel fatto che, come verrà mostrato nel seguito del paragrafo, il metodo basato sull'utilizzo di un test standard dà buoni risultati e risulta migliore in termini di efficienza rispetto al metodo degli intervalli di confidenza, solo nel caso in cui la condizione di costanza sia presente; nel caso contrario, quando cioè la suddetta condizione non possa essere assunta, l'applicazione di questo metodo potrebbe portare ad un rapido aumento dell'errore del primo tipo. La metodologia basata sugli intervalli di confidenza invece risulta essere più conservativa rispetto a questa problematica.

Sostanzialmente il problema è legato al fatto che il margine di non inferiorità non può mai essere determinato con certezza, e di conseguenza l'errore α del primo tipo non riesce a essere determinato con sicurezza. Riprendendo allora il discorso, anticipato nel paragrafo 2, sulla scelta del margine di non inferiorità, in generale si può affermare che questo margine dipende da una molteplicità di fattori clinici, e la sua determinazione risulta di conseguenza essere una questione alquanto complicata. Il significato di questo margine sta nel fatto, come abbiamo già detto, che spesso in un farmaco ha senso rinunciare ad una certa quantità di efficacia se questa può essere recuperata in termini di sicurezza, costi, facilità di somministrazione, etc..

I fattori che principalmente influenzano le decisioni relative al margine di non-inferiorità sono comunque l'ordine di grandezza, la variabilità e la costanza nel tempo dell'effetto del trattamento che nello studio viene preso come confronto. Come abbiamo già visto, spesso si usa definire il suddetto margine come percentuale, sufficientemente piccola, dell'effetto del trattamento di controllo, in modo da poter assicurare attraverso il test di non-inferiorità anche la superiorità del farmaco sperimentale sul placebo. Come abbiamo visto, si cerca di descrivere questa quantità come una funzione esplicita dell'effetto stesso del trattamento di controllo sul placebo, ma questo non è sempre possibile da fare: spesso accade che i valori registrati per il suddetto effetto nello studio in corso siano diversi dai valori storici riportati dalla letteratura, e questo rappresenta un vero problema in termini di affidabilità della stima del margine.

Quest'incertezza sulla determinazione del margine di non-inferiorità ha forti implicazioni, come abbiamo detto, sull'entità dell'errore α del I tipo. Di seguito verrà discussa la questione nel caso di utilizzo di intervalli di confidenza prima e

di test standard poi, prendendo in considerazione le ipotesi del test di non-inferiorità basate sul logaritmo degli effetti dei trattamenti introdotte nel secondo paragrafo[12]:

$$\begin{cases} H_0^{\downarrow}: & \log(T) - \log(C) \geq \delta \\ H_1^{\downarrow}: & \log(T) - \log(C) < \delta \end{cases}$$

dove, ricordiamo, gli effetti dei vari trattamenti sono intesi sotto forma di rischio relativo, e il margine di non-inferiorità δ è stabilito fissando una costante γ tale che:

$$\delta = (1-\gamma)[\log(P) - \log(C)]$$

Intervalli di confidenza:

Questo approccio prevede la costruzione di un intervallo di confidenza simmetrico ad un livello (1 - 2α) per la quantità log(T/C), e il confronto con la quantità scelta per δ, accettando l'ipotesi alternativa di non-inferiorità nel caso in cui il suddetto intervallo contenga interamente valori più piccoli di δ.

Quando il valore di δ è una costante fissa e nota è facile dimostrare che l'errore del I tipo, cioè di accettare l'ipotesi nulla $H_0^{\downarrow}$ quando essa è invece falsa, risulta al più pari ad α. In questo caso l'errore considerato viene calcolato semplicemente sui dati, relativamente al trattamento di controllo, disponibili dallo studio che si sta conducendo.

Quando invece la quantità δ deve essere stimata dai dati storici a disposizione, deve essere di conseguenza tenuto conto dell'incertezza legata alla stima che si sta utilizzando quando si va a calcolare l'entità dell'errore del I tipo, questo anche se è verificata la condizione di costanza nel tempo.

Infatti in questo caso, fissata la quantità γ, il margine di non-inferiorità δ verrà ottenuto tramite una stima basata sull'effetto del controllo, e cioè, nel nostro sistema di ipotesi, della quantità $[\log(P) - \log(C)]$; se la condizione di costanza nel tempo può essere ipotizzata vera, la suddetta quantità sarà stimabile attraverso una stima dell'analoga quantità relativa ai dati storici $[\log(P_0) - \log(C_0)]$ (stima che in seguito indicheremo con $\left[\log\left(\tilde{P}_0\right) - \log\left(\tilde{C}_0\right)\right]$),

[12] Per i riferimenti bibliografici vedere nota n. 6.

oppure, in un'ottica più conservativa, tramite l'estremo inferiore dell'intervallo di confidenza per questa quantità.

Vediamo cosa accade al valore effettivo dell'errore del I tipo, che chiameremo α^I, nei due casi, scegliendo prima come stima per l'effetto del trattamento di controllo l'alternativa più conservativa tra quelle menzionate sopra, e poi la semplice stima puntuale[13]:

- Scegliendo come stima per la quantità [log(P) – log(C)] l'estremo inferiore dell'intervallo di confidenza per la stima $\left[\log\left(\tilde{P}_0\right)-\log\left(\tilde{C}_0\right)\right]$, allora si avrà che, dato che l'intervallo di confidenza risultante per la quantità log(T) – log(C) è dato da:

  $$\log\left(\hat{T}\right)-\log\left(\hat{C}\right)\pm z_{1-\alpha}\sigma_{TC}$$

  e la stima per il margine di non-inferiorità, considerando come accennato prima l'estremo inferiore dell'intervallo di confidenza per $\left[\log\left(P_0\right)-\log\left(C_0\right)\right]$, è data da:

  $$\delta=(1-\gamma)\left[\log\left(\tilde{P}_0\right)-\log\left(\tilde{C}_0\right)-z_{1-\alpha}\sigma_{PC0}\right]$$

  dove σ_{TC} e σ_{PC0} sono gli errori standard delle rispettive stime, allora il metodo qui considerato rifiuterà l'ipotesi nulla H^I_0 in favore di quella alternativa quando:

  $$\log\left(\hat{T}\right)-\log\left(\hat{C}\right)+z_{1-\alpha}\sigma_{TC}<(1-\gamma)\left[\log\left(\tilde{P}_0\right)-\log\left(\tilde{C}_0\right)-z_{1-\alpha}\sigma_{PC0}\right]$$

  Si nota qui come giustamente l'incertezza legata alla determinazione del margine di non-inferiorità sia inclusa nella regola di decisione per il problema, e come di conseguenza il massimo errore del I tipo associato alla regione di rifiuto sia dato da:

$$\alpha^I=\Pr\left\{\log\left(\hat{T}\right)-\log\left(\hat{C}\right)+z_{1-\alpha}\sigma_{TC}<(1-\gamma)\left[\log\left(\tilde{P}_0\right)-\log\left(\tilde{C}_0\right)-z_{1-\alpha}\sigma_{PC0}\right]\right\}=$$

$$=\Pr\left\{\frac{\log\left(\hat{T}\right)-\log\left(\hat{C}\right)-(1-\gamma)\left[\log\left(\tilde{P}_0\right)-\log\left(\tilde{C}_0\right)\right]}{\sqrt{\left[\sigma^2_{TC}+(1-\gamma)^2\sigma^2_{PC0}\right]}}<-z_{1-\alpha}f\right\}=\Phi\left(-z_{1-\alpha}f\right)$$

[13] Il riferimento bibliografico di seguito riportato è relativo a tutte le formulazioni presentate in questa sezione: Hung H. M. J., Wang S., Tsong Y., Lawrence J., O'Neil R. T.; "Some fundamental issues with non-inferiority testing in active controlled trials", in *Statistics in Medicine*, (pp 217-218-219); 2003.

dove Φ rappresenta la funzione di ripartizione di una normale standard, e il parametro *f* è dato da:

$$f = [\sigma_{TC} + (1-\gamma)\sigma_{PC0}] / \left[\sigma_{TC}^2 + (1-\gamma)^2 \sigma_{PC0}^2\right]^{1/2}$$

ed è sempre maggiore di 1; Quindi dato che, se *f* > 1:

$$\alpha = \Phi(-z_{1-\alpha}) > \Phi(-z_{1-\alpha} f)$$

allora α^I, cioè l'effettivo errore del I tipo calcolato tenendo conto dell'incertezza legata alla stima del margine di non inferiorità, risulta sempre minore dell'α scelto all'inizio per costruire l'intervallo di confidenza per la decisione finale.

Possiamo allora concludere che il metodo degli intervalli di confidenza, che utilizza come stima per il margine di non-inferiorità il limite inferiore di un intervallo di confidenza per l'effetto del trattamento di controllo calcolato basandosi sui dati storici presenti in letteratura, sotto la condizione di costanza nel tempo, consente di mantenere basso l'effettivo errore del I tipo α^I, che risulta sempre minore di α.

- Scegliendo come stima per la quantità $[\log(P) - \log(C)]$ la semplice stima puntuale $\left[\log(\tilde{P}_0) - \log(\tilde{C}_0)\right]$, allora il metodo considerato rifiuterà l'ipotesi nulla H_0^I in favore di quella alternativa quando:

$$\log(\hat{T}) - \log(\hat{C}) + z_{1-\alpha}\sigma_{TC} < (1-\gamma)\left[\log(\tilde{P}_0) - \log(\tilde{C}_0)\right]$$

Anche in questo caso si ha inclusa nella decisione per il problema l'incertezza legata alla determinazione del margine di non-inferiorità, e di conseguenza, ripetendo i passaggi fatti al punto precedente, si ottiene:

$$\alpha^I = \Pr\left\{\log(\hat{T}) - \log(\hat{C}) + z_{1-\alpha}\sigma_{TC} < (1-\gamma)\left[\log(\tilde{P}_0) - \log(\tilde{C}_0)\right]\right\} =$$
$$= \Phi(-z_{1-\alpha} h)$$

dove il parametro *h* è dato da:

$$h = \sigma_{TC} / \left[\sigma_{TC}^2 + (1-\gamma)^2 \sigma_{PC0}^2\right]^{1/2}$$

ed è positivo e sempre minore di 1. Quindi dato che, se 0 < *h* < 1:

$$\alpha = \Phi(-z_{1-\alpha}) < \Phi(-z_{1-\alpha} h)$$

allora α^I, cioè l'effettivo errore del I tipo calcolato tenendo conto dell'incertezza legata alla stima del margine di non inferiorità, risulta maggiore dell'α scelto all'inizio per costruire l'intervallo di confidenza per

la decisione finale, e quindi la scelta meno conservativa potrebbe portare ad errori anche grossolani.

Visti i risultati descritti finora viene naturale la necessità di ricercare una definizione per il margine di non-inferiorità che riesca a produrre un errore α^{I} che sia esattamente uguale alla soglia prefissata α. La soluzione a questa questione viene fornita dalla seguente formulazione per il margine di non-inferiorità:

$$\delta^{*} = -z_{1-\alpha}\left\{\left[\sigma_{TC}^{2} + (1-\gamma)^{2}\sigma_{PC0}^{2}\right]^{1/2} - \sigma_{TC}\right\} + (1-\gamma)\left[\log\left(\tilde{P}_{0}\right) - \log\left(\tilde{C}_{0}\right)\right]$$

infatti la regione di rifiuto conseguente, data da:

$$\log\left(\hat{T}\right) - \log\left(\hat{C}\right) + z_{1-\alpha}\sigma_{TC} < \delta^{*}$$

ha un errore del I tipo associato pari esattamente a:

$$\alpha^{I} = \Pr\left\{\log\left(\hat{T}\right) - \log\left(\hat{C}\right) + z_{1-\alpha}\sigma_{TC} < \delta^{*}\right\} =$$

$$= \Pr\left\{\frac{\log\left(\hat{T}\right) - \log\left(\hat{C}\right) - (1-\gamma)\left[\log\left(\tilde{P}_{0}\right) - \log\left(\tilde{C}_{0}\right)\right]}{\sqrt{\left[\sigma_{TC}^{2} + (1-\gamma)^{2}\sigma_{PC0}^{2}\right]}} < -z_{1-\alpha}\right\} = \Phi\left(-z_{1-\alpha}\right) = \alpha$$

Analizzando questa nuova formulazione per il margine di non-inferiorità ci accorgiamo che esso è funzione della quantità σ_{TC}, e cioè dell'errore standard dell'effetto del trattamento sperimentale rapportato a quello di controllo, quantità che dipende solo dall'esperimento che si sta conducendo, e non dai dati storici come sarebbe più corretto; in particolar modo qualche perplessità può sorgere se si pensa che il suddetto errore standard è direttamente legato all'ampiezza del campione analizzato, e che di conseguenza il margine di non inferiorità, che dovrebbe essere determinato a priori ad esperimento non ancora cominciato, necessita invece per la sua determinazione della numerosità campionaria, la cui ampiezza dovrebbe invece essere pianificata in un secondo momento, violando così i principi teorici per questo tipo di studi, discussi nel primo capitolo.

Operativamente comunque la formulazione di δ^{*} ha le seguenti caratteristiche:

- è una funzione direttamente proporzionale alla quantità σ_{TC}, in quanto la sua derivata rispetto a questa variabile risulta sempre positiva:

$$\left[\delta^{*}\right]' = -z_{1-\alpha}\left[\frac{\sigma_{TC}}{\sqrt{\left[\sigma_{TC}^{2} + (1-\gamma)^{2}\sigma_{PC0}^{2}\right]}} - 1\right] > 0$$

e quindi è anche una quantità che diminuisce all'aumentare della numerosità campionaria, visto che l'errore standard sopra considerato è inversamente proporzionale, come è noto, alla radice quadrata dell'ampiezza campionaria stessa.

- per $\sigma_{TC} \to \infty$, e quindi per numerosità campionarie basse, la quantità $\left\{\left[\sigma_{TC}^2 + (1-\gamma)^2 \sigma_{PC0}^2\right]^{1/2} - \sigma_{TC}\right\} \to 0$, e di conseguenza il margine associato δ^* tende a coincidere con quello considerato nel secondo caso sopra analizzato, in cui si stima la quantità [log(P) – log(C)] con la semplice stima puntuale $\left[\log\left(\tilde{P}_0\right) - \log\left(\tilde{C}_0\right)\right]$, caso in cui, come abbiamo dimostrato, l'errore del I tipo associato alla regione di rifiuto risulta essere maggiore del valore scelto per α.
- per $\sigma_{TC} \to 0$, e quindi per numerosità campionarie alte, la quantità $\left\{\left[\sigma_{TC}^2 + (1-\gamma)^2 \sigma_{PC0}^2\right]^{1/2} - \sigma_{TC}\right\} \to (1-\gamma)\sigma_{PC0}$, e di conseguenza il margine associato tende a coincidere con quello considerato nel primo caso sopra analizzato, in cui si stima la quantità [log(P) – log(C)] con l'estremo inferiore dell'intervallo di confidenza per la stima $\left[\log\left(\tilde{P}_0\right) - \log\left(\tilde{C}_0\right)\right]$, caso in cui, come abbiamo dimostrato, l'errore del I tipo associato alla regione di rifiuto risulta essere minore del valore scelto per α.

Le osservazioni fatte finora portano ad una conclusione molto significativa: per campioni molto numerosi il metodo di scelta basato sugli intervalli di confidenza richiede la scelta di un margine del non-inferiorità più piccolo per garantire il controllo dell'errore del I tipo associato, esattamente pari al valore scelto per α.

Quindi risolvere il problema della non-inferiorità utilizzando l'approccio basato sugli intervalli di confidenza nel caso in cui il margine di non inferiorità non sia una quantità fissa stabilita con certezza, ma la sua determinazione si basi su una stima dell'effetto del farmaco di controllo sul placebo derivante dai dati storici presenti in letteratura, diventa una procedura delicata, in cui bisogna prestare grande attenzione alle condizioni in cui si porta avanti lo studio, e in particolare alla numerosità campionaria che si ha a disposizione, in quanto quest'ultima ha una forte influenza sull'entità dell'errore di I specie legato alla regione di rifiuto dell'ipotesi nulla da testare.

Test standard:

Il metodo di soluzione basato sull'utilizzo di una statistica test da confrontare con una distribuzione teorica può essere utilizzato con ottimi risultati per risolvere i problemi di non-inferiorità quando si è autorizzati ad assumere la condizione di costanza nel tempo dell'effetto del trattamento di controllo sul placebo.

In questo caso infatti si possono testare le ipotesi

$$\begin{cases} H_0^{|}: & \log(T) - \log(C) \geq \delta \\ H_1^{|}: & \log(T) - \log(C) < \delta \end{cases}$$

attraverso un test, chiamato "preservation test" così definito[14]:

$$Z_{pv} = \frac{\log(\hat{T}) - \log(\hat{C}) - (1-\gamma)\left[\log(\tilde{P}_0) - \log(\tilde{C}_0)\right]}{\sqrt{\left[\sigma_{TC}^2 + (1-\gamma)^2 \sigma_{PC0}^2\right]}}$$

caratterizzato dalla seguente regione di rifiuto dell'ipotesi nulla:

$$Z_{pv} < -z_{1-\alpha}$$

In questo caso l'effettivo errore di I specie $\alpha^{|}$ associato alla regione di rifiuto sopra descritta risulta essere esattamente pari ad α.

È interessante notare come questo criterio di rifiuto sia matematicamente equivalente a quello che si utilizza con il metodo degli intervalli di confidenza scegliendo per il margine di non-inferiorità il valore δ^* definito nel paragrafo precedente. Bisogna però specificare che le interpretazioni del margine di non-inferiorità nei due metodi sono differenti: infatti in questo caso, come definito all'inizio di questo paragrafo, il margine che si considera è dato da:

$$\delta = (1-\gamma)\left[\log(P) - \log(C)\right]$$

inteso come parametro fissato a priori, e non più allora stimato in modo da ottenere i risultati desiderati.

Questo test però riesce a tenere controllato l'errore di I specie solamente nel caso in cui sussista la condizione di costanza nel tempo: infatti solo in questo modo si è autorizzati a considerare ragionevole la scelta del margine di non-inferiorità come definito sopra, cioè basandosi solo sui dati relativi allo studio in

[14] Il riferimento bibliografico di seguito riportato è relativo a tutte le formulazioni presentate in questa sezione: Hung H. M. J., Wang S., Tsong Y., Lawrence J., O'Neil R. T.; "Some fundamental issues with non-inferiority testing in active controlled trials", in *Statistics in Medicine*, (pp 219); 2003.

corso. Nel caso in cui invece l'effetto del trattamento di controllo sia diverso nel tempo, e in particolare quando accade che esso registri valori più bassi di efficacia nell'esperimento corrente rispetto ai dati storici riscontrabili in letteratura, succede che l'effettivo errore di I specie α^I eccede il valore di α.

In conclusione, mettendo a confronto i due metodi di decisione fin qui analizzati, per quanto riguarda la questione della variabilità dell'errore di I specie, possiamo dire che:

- in caso di costanza dell'effetto del trattamento di controllo nel tempo, il metodo del test standard risulta migliore del metodo degli intervalli di confidenza che utilizza come stima per la quantità [log(P) – log(C)] la semplice stima puntuale $\left[\log\left(\tilde{P}_0\right) - \log\left(\tilde{C}_0\right)\right]$, in quanto l'entità dell'errore α^I risulta minore nel caso dell'utilizzo della statistica test;
- sempre in caso di validità della condizione di costanza, il metodo del test standard e quello degli intervalli di confidenza che utilizza come valore del margine di non-inferiorità esattamente δ^*, coincidono dal punto di vista considerato, in quanto in entrambi i casi l'errore α^I coincide esattamente con α;
- nel caso in cui sia violata la condizione di costanza, il metodo degli intervalli di confidenza che utilizza come stima per la quantità [log(P) – log(C)] l'estremo inferiore dell'intervallo di confidenza per la stima $\left[\log\left(\tilde{P}_0\right) - \log\left(\tilde{C}_0\right)\right]$ risulta migliore del metodo del test standard, in quanto meno sensibile all'assenza della suddetta condizione.

1.3.7. Il metodo TACT

Come già accennato in sede di introduzione a questo capitolo, il metodo TACT è stato sviluppato al fine di gestire al meglio il problema del controllo dell'errore del primo tipo in relazione all'esistenza della condizione di costanza nel tempo di cui si è trattato nei capitoli precedenti, sfruttando al meglio i punti di forza delle due metodologie normalmente utilizzate (intervalli di confidenza e test standard), in modo tale da supplire con una le debolezze dell'altra e viceversa.

Di seguito verrà descritto il metodo in tutti i suoi passi, ne verranno poi mostrate le performance in un confronto con le altre due metodologie e infine verranno tratte le conclusioni in merito agli argomenti esposti.

Supponiamo di voler testare l'efficacia di un farmaco sperimentale in termini di incidenza di un determinato evento in un determinato periodo di tempo. Verranno adottate le stesse notazioni, per indicare gli effetti dei vari trattamenti nell'esperimento in corso e relativamente ai dati storici, utilizzate anche nei capitoli precedenti (T, C, P, P_0, C_0).

Supponiamo che l'esperimento che si sta conducendo non disponga di dati relativi al placebo e assumiamo inoltre che l'effetto del placebo sia invariato rispetto a quello riscontrabile nei dati storici. Allora possiamo formulare il seguente modello per descrivere l'effetto del trattamento, in termini di logaritmo del rischio relativo[15]:

$$\log(incidenza - evento) = \mu + \beta X_{C0} + \gamma X_C + \xi X_T$$

dove $\mu = \log(P) = \log(P_0)$ è l'effetto del placebo;

$\beta = \log(C_0 / P)$ è l'effetto del farmaco di controllo rispettivamente all'effetto del placebo nei dati storici;

$\gamma = \log(C/P)$ è l'effetto del farmaco di controllo rispettivamente all'effetto del placebo nell'esperimento in atto;

$\xi = \log(T/P)$ è l'effetto del farmaco sperimentale rispettivamente all'effetto del placebo;

X_{C0}, X_C e X_T sono le variabili indicatrici associate ai vari trattamenti.

Relativamente al modello sopra descritto possono venire formulate le seguenti ipotesi di interesse: la presenza di una certa efficacia del farmaco sperimentale viene formalizzata dal seguente sistema:

$$\begin{cases} H_0: & \xi \geq 0 \\ H_1: & \xi < 0 \end{cases}$$

mentre l'ipotesi di non inferiorità, e quindi di capacità del farmaco sperimentale di preservare il 100λ per cento dell'effetto del farmaco di controllo viene invece descritta dal sistema seguente:

[15] Il riferimento bibliografico di seguito riportato è relativo a tutte le formulazioni presentate in questa sezione: Wang S., Hung H. M. J.; "TACT method for non-inferiority testing in active controlled trials", in *Statistics in Medicine*, (pp 229-230-231); 2003.

$$\begin{cases} H_{0\lambda}: & \xi - \gamma \geq \delta \\ H_{1\lambda}: & \xi - \gamma < \delta \end{cases}$$

con $\delta = -(1-\lambda)\gamma$ il margine di non-inferiorità.

Partendo dal modello sopra descritto, il metodo TACT si sviluppa su due fasi distinte:

i. vi è una prima fase, nella quale si accerta la superiorità del farmaco di controllo rispetto al placebo studiando i dati storici a disposizione: una volta accertata la suddetta superiorità si può procedere con le fasi successive;
ii. la seconda fase consiste in una analisi congiunta dei dati storici ottenuti dalla ricerca nella letteratura e di quelli relativi all'esperimento in corso, che si compone a sua volta di due step:
    - si va innanzitutto a confrontare i valori registrati per l'effetto del farmaco di controllo nei dati storici e nell'esperimento corrente, per stabilire se le differenze siano troppo elevate: in questo caso (cioè se γ >> β, e cioè se viene a mancare la condizione di costanza nel tempo) i due metodi tradizionali di soluzione, quello basato sugli intervalli di confidenza e quello sul test standard, diventano poco affidabili a causa del possibile aumento dell'errore del I tipo, come descritto nel paragrafo precedente. Questa analisi viene condotta tramite il confronto di una statistica test che stimi la differenza tra gli effetti fatti registrare dal trattamento di controllo nei diversi tempi, dati storici ed esperimento corrente, con un valore opportuno U_t scelto in base a considerazione cliniche e statistiche; lo studio deve essere ripetuto considerando non solo il confronto tra il tempo 0 e il tempo 1 (l'esperimento corrente), ma deve venire condotto per una serie di tempi intermedi, confrontando cioè i valori registrati al tempo 0 con quelli registrati ad un tempo t con 0 < t < 1; il test in questione è definito nel modo seguente:

      $$Z_{Ct} = \frac{\log\left(\hat{C}_t\right) - \log\left(\hat{C}_0\right)}{\sigma}$$

      dove $\hat{C}_t$ e $\hat{C}_0$ sono le stime per i valori relativi al trattamento di controllo rispettivamente nell'esperimento al determinato istante t e nel tempo 0, e σ è l'errore standard della quantità al numeratore, e la decisione di rifiuto dell'ipotesi di costanza dell'effetto nel tempo viene presa quando:

$$Z_{Ct} > U_t$$

Se viene verificata questa situazione allora conviene sospendere l'esperimento in quanto l'obiettivo di stabilire la non-inferiorità del trattamento sperimentale sul farmaco di controllo, e quindi indirettamente, per come abbiamo definito nei paragrafi precedenti il margine di non-inferiorità quale percentuale dell'effetto del trattamento di controllo sul placebo, anche la superiorità del farmaco testato sul placebo stesso;

- si procede con lo studio se le analisi descritte sopra portano ad affermare che l'effetto del trattamento di controllo è costante nel tempo (cioè se viene rifiutata l'ipotesi che γ >> β);in questo secondo step si va a considerare il valore della statistica Z_{C1}, cioè il valore del test calcolato al tempo t = 1 e quindi per quanto riguarda l'esperimento corrente, per decidere quale metodo di decisione adottare per una maggior affidabilità dei risultati. Facendo riferimento a quanto concluso nel paragrafo precedente, si potranno prendere tre diverse decisioni in base al valore registrato per il test Z_{C1}: i) se il test assume un valore grande ($Z_{C1} > U$) allora conviene abbandonare lo studio per i motivi già presentati poco sopra, mentre ii) se il test assume valore bassi ($Z_{C1} < L$) si può decidere di utilizzare per il test di non inferiorità il metodo del test standard se il valore di Z_{C1} risulta particolarmente ridotto, oppure il metodo degli intervalli di confidenza se il valore di Z_{C1} denota comunque una certa differenza di effetto del trattamento di controllo, in un'ottica più conservativa dal punto di vista del controllo dell'errore del I tipo.

Per quanto riguarda il metodo del test tradizionale la statistica da usare sarà il "preservation test" descritto nel paragrafo precedente:

$$Z_{pv} = \frac{\log\left(\hat{T}\right) - \log\left(\hat{C}\right) - (1-\gamma)\left[\log\left(\tilde{P}_0\right) - \log\left(\tilde{C}_0\right)\right]}{\sqrt{\left[\sigma_{TC}^2 + (1-\gamma)^2 \sigma_{PC0}^2\right]}}$$

caratterizzato dalla seguente regione di rifiuto dell'ipotesi nulla:

$$Z_{pv} < -z_{1-\alpha}$$

Per il metodo degli intervalli di confidenza dovrà essere considerata la regola di decisione più conservativa, anch'essa descritta in modo più approfondito nel paragrafo precedente:

$$\log\left(\hat{T}\right) - \log\left(\hat{C}\right) + z_{1-\alpha}\sigma_{TC} < (1-\gamma)\left[\log\left(\tilde{P}_0\right) - \log\left(\tilde{C}_0\right) - z_{1-\alpha}\sigma_{PC0}\right]$$

1.3.8. Le performance del metodo TACT

Sono stati condotti diversi studi di simulazione per dimostrare l'efficienza del metodo TACT rapportato agli altri due metodi tradizionali di decisione per il problema della non-inferiorità. Di seguito verranno riportate le conclusioni principali a cui si è giunti per dare un quadro generale delle caratteristiche delle tre metodologie messe a confronto nelle diverse situazioni che si possono presentare durante uno esperimento clinico:

- nel caso in cui la condizione di costanza nel tempo sia fortemente violata (e cioè nel caso in cui γ >> β) il metodo del test standard risulta essere una soluzione poco affidabile, come il metodo degli intervalli di confidenza, per cui in entrambi i casi risulta necessario abbandonare lo studio. Il metodo TACT, per sua stessa costruzione, è in grado di riconoscere il problema e di dare già indicazioni sul comportamento da assumere;
- nel caso in cui la violazione della condizione di costanza risulti poco chiara, optare per uno dei due metodi tradizionali può portare ad errori anche grossolani in quanto in entrambi i casi vi è una crescita dell'errore α sopra il target prestabilito (crescita, come già descritto, più veloce nell'utilizzo del test standard piuttosto che nell'intervallo di confidenza). In questo tipo di situazione il metodo TACT risulta tenere un errore del I tipo più basso, risulta essere quindi più conservativo;
- per quanto riguarda i casi in cui può essere assunta la condizione di costanza nel tempo, si può fare un confronto interessante anche tra la potenza delle tre metodologie: si dimostra che il metodo TACT presenta una potenza che si avvicina molto a quella del test standard, che come abbiamo visto è il migliore tra i due approcci standard sotto la suddetta condizione. Il piccolo guadagno di potenza che rimane a favore del test

standard può essere attribuito se non altro alla maggiore velocità di applicazione di quest'ultimo rispetto a metodo TACT. In generale il metodo degli intervalli di confidenza risulta il meno potente dei tre, in particolar modo l'approccio che basa la stima dell'effetto del controllo sull'estremo inferiore dell'intervallo di confidenza (come descritto nel paragrafo 1.3.6.);

Infine possiamo dire che la forza del metodo TACT sta essenzialmente nell'analisi preliminare che viene premessa all'applicazione dei metodi di decisione per il problema della non-inferiorità: infatti questo passaggio permette di preservare l'errore del I tipo dall'incremento che invece esso subisce in assenza della condizione di costanza nel tempo negli altri due metodi standard. Inoltre esso può godere, in caso di presenza della condizione di costanza, della potenza e dei pregi del metodo del test standard, oppure delle proprietà conservative del metodo degli intervalli di confidenza.

1.3.9. I test basati sui ranghi

Mi è parso inevitabile dedicare un paragrafo alla descrizione di alcune di queste metodologie, in quanto come noto nella realtà applicativa molto più spesso occorre ricorrere ad esse piuttosto che a soluzioni parametriche, le cui assunzioni di base possono risultare troppo forti e pertanto molto meno spesso applicabili. Verrà mostrato in seguito come questi test non parametrici rappresentano un'efficiente alternativa alle versioni parametriche quando non vi sono le condizioni per applicare queste ultime, e raggiungono comunque livelli di efficienza molto simili a questi quando invece vi sono le suddette condizioni. Inoltre questi stimatori non richiedono trasformazioni delle variabili oggetto di studio per ricercare la normalità e sono meno sensibili ai valori anomali rispetto alle loro versioni parametriche, aspetti che aumentano il valore delle motivazioni a favore di questi test.

Nel presente paragrafo verranno descritte le procedure non-parametriche utilizzate per risolvere il problema della non-inferiorità nei frequenti casi in cui le assunzioni di base per l'approccio parametrico non possano essere sostenute. Verrà introdotto il problema nell'ambito di studi clinici condotti su più centri

differenti, in modo da definire le assunzioni di base in un'ottica generale; si passerà poi alla definizione della metodologia di soluzione nello specifico caso di due trattamenti (il farmaco sperimentale e il trattamento di controllo) somministrati in un solo centro.

Prendiamo allora in considerazione il caso generale dell'analisi parallela su più centri clinici: qui la scelta dei centri da includere nel campione deve essere effettuata in modo casuale, e all'interno di ogni centro la somministrazione di trattamento sperimentale, trattamento di controllo e, se presente, placebo ai vari pazienti viene fatta ancora in modo casuale. Può accadere allora che i vari centri risultino eterogenei per quanto riguarda la numerosità di soggetti analizzati o la variabilità tra di essi. Appare chiaro allora come in una situazione come questa sia facile che vengano a mancare le condizioni base per adottare metodi di decisione che usano approcci parametrici, e convenga allora utilizzare metodologie di tipo non parametrico.

Le assunzioni di base per l'applicazione di questi test sono, oltre alla scelta, già accennata, casuale dei centri da includere nello studio, la numerosità interna a ciascun centro piuttosto elevata e la presenza di un solo rilevatore all'interno di ciascun centro per garantire l'omogeneità delle misurazioni.

Assumendo: t trattamenti;

c centri;

n_{ij} replicazioni dell'i-esimo trattamento (con $i = 1,\ldots,t$) nel j -esimo centro (con $j = 1,\ldots,c$);

$n_{.j} = \sum_{j=1}^{t} n_{ij}$ osservazioni totali che vengono fatte in ogni centro;

Y_{ijl} la singola osservazione dalla l -esima replicazione dell'i-esimo trattamento nel j -esimo centro;

allora il modello generale[16] che descrive l'esperimento è dato da:

$$Y_{ijl} = \theta_i + \beta_j + e_{ijl} \qquad \text{con} \quad i = 1,\ldots,t; \quad j = 1,\ldots,c; \quad l = 1,\ldots, n_{ij}$$

dove θ_i è l'effetto (media o mediana) dell'i-esimo trattamento;

β_j è l'effetto del j –esimo centro, e viene assunto che i β_j (per $j = 1,\ldots,c$) sono variabili casuali indipendenti e identicamente distribuite di media 0 e varianza σ_β^2;

[16] Rashid M. M.; "Rank-based tests for non-inferiority and equivalence hypotesis in multi-centre clinical trials using mixed models", in *Statistics in Medicine*, (pp 293-294); 2003.

e_{ijl} è l'errore casuale associato alla singola osservazione, e viene assunto che gli e_{ijl} (per i = 1,…,t e j = 1,…,c) sono variabili casuali indipendenti e identicamente distribuite di media 0 e varianza σ_e^2;

β_j e e_{ijl} sono indipendenti.

Sotto l'assunzione che l'effetto del trattamento sia simile nei diversi centri, e quindi che l'interazione esistente tra tipo di trattamento e centro sia trascurabile, e considerando la nuova variabile $\varepsilon_{ijl} = \beta_j + e_{ijl}$, si interpreta allora $Y_j = (y_{1j},\ldots, y_{ij},\ldots, y_{tj})^T$ come il vettore di dimensioni $n_{\cdot j}$ x 1 delle singole osservazioni per il j-esimo centro, dove $y_{ij} = (y_{ij1},\ldots, y_{ijl},\ldots, y_{ijn_{ij}})^T$ è il vettore delle singole osservazioni del gruppo trattato con il trattamento i nel centro j; analogamente si avrà che ε_j sarà il vettore degli errori del modello considerato per il j-esimo centro.

Come noto l'inferenza parametrica assume che gli ε_j (per j = 1,…,c) siano variabili casuali normali multivariate con media 0 e matrice di covarianza data da $\sigma^2[(1-\rho)\ I_{n_{\cdot j} X n_{\cdot j}} + \rho\ 1_{n_{\cdot j}}\ 1^T_{n_{\cdot j}}]$, dove $\sigma^2 = \mathrm{var}(\varepsilon_{ijl}) = \sigma_\beta^2 + \sigma_e^2$, il coefficiente ρ è dato da $\sigma_\beta^2 / (\sigma_\beta^2 + \sigma_e^2)$, $I_{n_{\cdot j} X n_{\cdot j}}$ è la matrice identica di ordine $n_{\cdot j}$ e $1_{n_{\cdot j}}$ è un vettore di unità di dimensioni $n_{\cdot j}$ x 1. Spesso però nei casi reali questo tipo di assunzione non riesce ad essere giustificata, e questo porta allora alla necessità, come dicevamo sopra, di adottare metodi di soluzione alternativi: si introdurrà di seguito una metodologia basata sulla minimizzazione della somma delle funzioni di dispersione di Jaeckel basate sui ranghi di residui, in modo da ottenere degli stimatori per gli θ_i.

Nel caso generale di t trattamenti somministrati su $n_{\cdot j}$ soggetti in ciascuno dei c centri (con j = 1,…,c), indicato con $\theta = (\theta_1,\ldots, \theta_t)$ il vettore degli effetti dei trattamenti, la funzione di dispersione di Jaeckel per il j-esimo centro basata sui punteggi di Wilcoxon è definita nel modo seguente[17]:

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

dove $a(W_{ijl}) = (n_{\cdot j} + 1)^{-1} R(W_{ijl}) - \frac{1}{2}$;

$W_{ijl} = Y_{ijl} - \theta_i$;

[17] Rashid M. M.; "Rank-based tests for non-inferiority and equivalence hypotesis in multi-centre clinical trials using mixed models", in *Statistics in Medicine*, (pp 295-296); 2003.

$R(W_{ijl})$ è il rango del residuo W_{ijl} corrispondente alla l-esima replicazione dell'i-esimo trattamento nel j-esimo centro.

È da notare nella formulazione sopra proposta che vengono utilizzati i pesi $(n_{.j}+1)^{-1}$ allo scopo di diminuire lo squilibrio di peso altrimenti fortemente esistente tra i centri più grandi e quelli invece aventi meno soggetti trattati. Inoltre $D_j(\theta)$ è una funzione lineare dei residui e questo favorisce una minore influenza dei valori anomali nella stima finale.

La funzione combinata di dispersione è definita allora come segue:

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

ed è anch'essa basata sui ranghi calcolati tra i vari centri.

Si dimostra che $D_j(\theta)$ e $D(\theta)$ sono entrambe funzioni non negative, continue e convesse in θ, in particolare si noti che per $t = 2$, $D(\theta)$ può essere ridotta ad una funzione di un solo parametro, infatti assunti come veri valori per gli effetti dei due trattamenti $\theta_2 = \mu$ e $\theta_1 = \mu + \delta$, si ha che $D(\theta_1, \theta_2) = D(\mu + \delta, \mu) = \sum_{j=1}^{c} D_j(\mu + \delta, \mu) = \sum_{j=1}^{c} D_j(\delta, 0) = D(\delta, 0)$ dove quindi $\delta = \theta_1 - \theta_2$.

Lo stimatore $\hat{\theta}$ basato sui ranghi per il vettore degli effetti dei trattamenti θ sarà determinato minimizzando la funzione di dispersione $D(\theta)$, quindi nel caso sopra descritto di due trattamenti si otterrà un valore $\hat{\delta}$. Si dimostra inoltre che lo stimatore $\hat{\delta}$ così trovato si distribuisce normalmente.

Nel caso specifico di interesse, due trattamenti somministrati a due gruppi di pazienti (t = 2) all'interno di un unico centro ($c = 1$), lo stimatore $\hat{\delta}$ corrisponde allo stimatore di Hodges-Lehmann dello shift $\delta = \theta_1 - \theta_2$ basato sul test sui ranghi di Wilcoxon-Mann-Whitney.

Per quanto riguarda il problema della non-inferiorità esso può essere riformulato, per rimanere in linea con la simbologia utilizzata nel presente paragrafo, nel modo seguente[18]:

$$\begin{cases} H_0: & \theta_1 \le \theta_2 - \delta_0 \\ H_1: & \theta_1 > \theta_2 - \delta_0 \end{cases}$$

[18] Rashid M. M.; "Rank-based tests for non-inferiority and equivalence hypotesis in multi-centre clinical trials using mixed models", in *Statistics in Medicine*, (pp 302-303); 2003.

dove θ_1 rappresenta il vero e sconosciuto effetto del trattamento sperimentale, θ_2 quello del trattamento di controllo e δ_0 è il margine di non-inferiorità stabilito a priori come descritto nei paragrafi precedenti. Anche qui, analogamente alle formulazioni precedenti, l'ipotesi nulla descrive il caso in cui il trattamento sperimentale risulti inferiore in termini di efficacia rispetto a quello scelto come controllo, mentre l'ipotesi alternativa viene accettata nel caso in cui il farmaco sperimentale possa essere considerato non inferiore al controllo meno il margine di non-inferiorità δ_0.

L'approccio non-parametrico basato sui ranghi si basa sulla seguente statistica test:

$$Z^* = \left[\hat{\delta} - \left(-\delta_0\right)\right] / SE_{\hat{\delta}}$$

e la regola di decisione, data la normalità dello stimatore $\hat{\delta}$, rifiuta l'ipotesi nulla nel caso in cui:

$$Z^* < -z_{1-\alpha}$$

Questa regola di decisione equivale peraltro a:

$$\hat{\delta} - z_{1-\alpha} SE_{\hat{\delta}} > -\delta_0$$

e quindi al confronto tra il limite inferiore dell'intervallo di confidenza al livello 100(1 - 2α) per il vero valore dello shift δ e il margine di non-inferiorità, metodo di decisione, come già accennato, spesso preferito dal personale medico perché maggiormente intuitivo.

Concludiamo la presente trattazione accennando ai vantaggi a cui porta l'approccio non-parametrico fin qui descritto, argomentando quindi quanto detto poco sopra a proposito dell'efficienza di questo tipo di test in relazione ai classici test parametrici: il test basato sui ranghi risulta più robusto rispetto alla soluzione parametrica in quanto, come mostrato, esso rimane meno sensibile alla presenza di valori anomali; inoltre non vi sono assunzioni di normalità da fare nel caso di approccio non parametrico, e questo aumenta le reali situazioni di applicazione e permette di risolvere studi che altrimenti con l'approccio parametrico dovrebbero essere abbandonati.

1.3.10. Studio di simulazione

Quest'ultimo paragrafo sarà dedicato alla descrizione dei risultati di uno studio di simulazione, condotto tramite l'uso del software R. Nello studio sono state considerate le ipotesi del test standard di Blackwelder basato sulla semplice differenza tra i trattamenti di controllo e sperimentale, che ricordiamo essere definito come segue:

$$\begin{cases} H_0: & \mu_C - \mu_T \geq \delta \\ H_1: & \mu_C - \mu_T < \delta \end{cases}$$

dove: μ_C = media della variabile indicatrice dell'efficacia del farmaco di controllo,

μ_T = media della variabile indicatrice dell'efficacia del farmaco sperimentale,

δ = margine di non inferiorità, definito >0

Come statistica test è stato utilizzato il test *t* di Student, per simmetria di distribuzione e semplicità di significato.

Per rimanere fedele all'iter di lavoro fin qui descritto, sono partita con la simulazione dei dati storici: ho assunto che la popolazione trattata con placebo si distribuisse come una variabile normale di media 0 e varianza unitaria; la popolazione trattata con il farmaco di controllo è stato invece supposto distribuirsi sempre normalmente e con varianza pari a 1, ma con media questa volta anch'essa uguale ad 1. Sono stati allora creati ad ogni replicazione (per un totale di 10000 replicazioni) due campioni indipendenti, di ampiezza uguale tra loro e pari a 50, dalle popolazioni sopra descritte; per ogni coppia di campioni è stato stimato lo shift della media delle due popolazioni tramite la differenza tra le due medie campionarie e infine è stata considerata come stima finale dell'effetto del trattamento di controllo sul placebo la media delle 10000 stime ottenute nelle altrettante replicazioni, che è risultata essere pari a $stima(C_0 - P_0) = 1.003118$. Di conseguenza allora è stato determinato il margine di non-inferiorità come percentuale di quest'ultima quantità stimata, come descritto nel corso del capitolo: ho ritenuto accettabile, e verosimilmente apprezzabile anche dallo staff medico dell'ipotetico studio in corso, richiedere al trattamento sperimentale di mantenere almeno l'80% dell'effetto del farmaco di

controllo sul placebo, e di conseguenza ho calcolato il margine di non inferiorità come il 20% del suddetto effetto, ottenendo $\delta = 0.2006235$.

Le ipotesi del test analizzato possono allora essere riscritte in questo caso nel modo seguente:

$$\begin{cases} H_0: & \mu_C - \mu_T \geq 0.2006235 \\ H_1: & \mu_C - \mu_T < 0.2006235 \end{cases}$$

A questo punto sono state simulate 33 coppie di campioni indipendenti, in cui il primo proveniente sempre dalla popolazione trattata con il farmaco di controllo come definita all'inizio del paragrafo, e il secondo proveniente dalla popolazione supposta trattata con il farmaco sperimentale, distribuita sempre normalmente, con varianza unitaria, ma per ogni coppia con media diversa, per simulare shift differenti tra i due trattamenti.

Per ogni coppia sono state fatte 10000 replicazioni, e sono state calcolate le probabilità di accettare l'ipotesi alternativa H_1 per 33 valori diversi del vero shift tra le medie delle due popolazioni, ottenendo i risultati riportati in tabella 3:

id	Shift	Probabilità
1	-1,00	1
2	-0,90	1
3	-0,80	0,9993
4	-0,70	0,9976
5	-0,60	0,9896
6	-0,50	0,9634
7	-0,45	0,9450
8	-0,40	0,9030
9	-0,35	0,8599
10	-0,30	0,8009
11	-0,25	0,7244
12	-0,20	0,6405
13	-0,15	0,5330
14	-0,10	0,4407
15	-0,05	0,3502
16	0,00	0,2584
17	0,05	0,1839
18	0,10	0,1289
19	0,15	0,0786
20	0,16	0,0786
21	0,17	0,0692
22	0,18	0,0605
23	0,19	0,0605
24	0,20	0,0492
25	0,2006235	0,0461
26	0,21	0,0485
27	0,22	0,0406
28	0,23	0,0386
29	0,24	0,0348
30	0,25	0,0265
31	0,30	0,0170
32	0,40	0,0050
33	0,50	0,0011
34	0,60	0,0002
35	0,70	0
36	0,80	0
37	0,90	0
38	1,00	0

Tabella 3: probabilità di accettazione dell'ipotesi alternativa H_1 per diversi valori della vera differenza (shift) della due medie degli effetti dei due trattamenti, di controllo e sperimentale; le suddette probabilità sono state calcolate come numero totale dei p-value inferiori a valore scelto per α=0.05 sul totale dei 10000 p-value calcolati per le altrettante replicazioni effettuate nello studio.

Osserviamo i casi in cui è vera l'ipotesi alternativa H_1, cioè quando la differenza tra l'effetto del trattamento di controllo e quello sperimentale è minore di $\delta = 0.2006235$: in questo caso l'ipotesi alternativa stessa presenta una probabilità di essere accettata, cioè il test presenta un p-value inferiore al valore scelto per α, che è quasi sempre (ad eccezione del caso dello shift pari a 0,20) maggiore di α, mentre questa probabilità scende sotto il livello di confidenza del test quando diventa vera l'ipotesi nulla H_0, cioè quando, casi riportati nell'ultima parte della tabella, lo shift tra gli effetti dei due trattamenti è maggiore del margine di non-inferiorità prescelto.

La non distorsione dimostrata dalla tabella appena analizzata può essere vista anche nel grafico riportato in figura 1, dove semplicemente vengono riportati sull'asse delle ascisse i valori simulati per la vera differenza tra i due trattamenti studiati, e sulle ordinate si trova la probabilità di accettazione dell'ipotesi alternativa H_1:

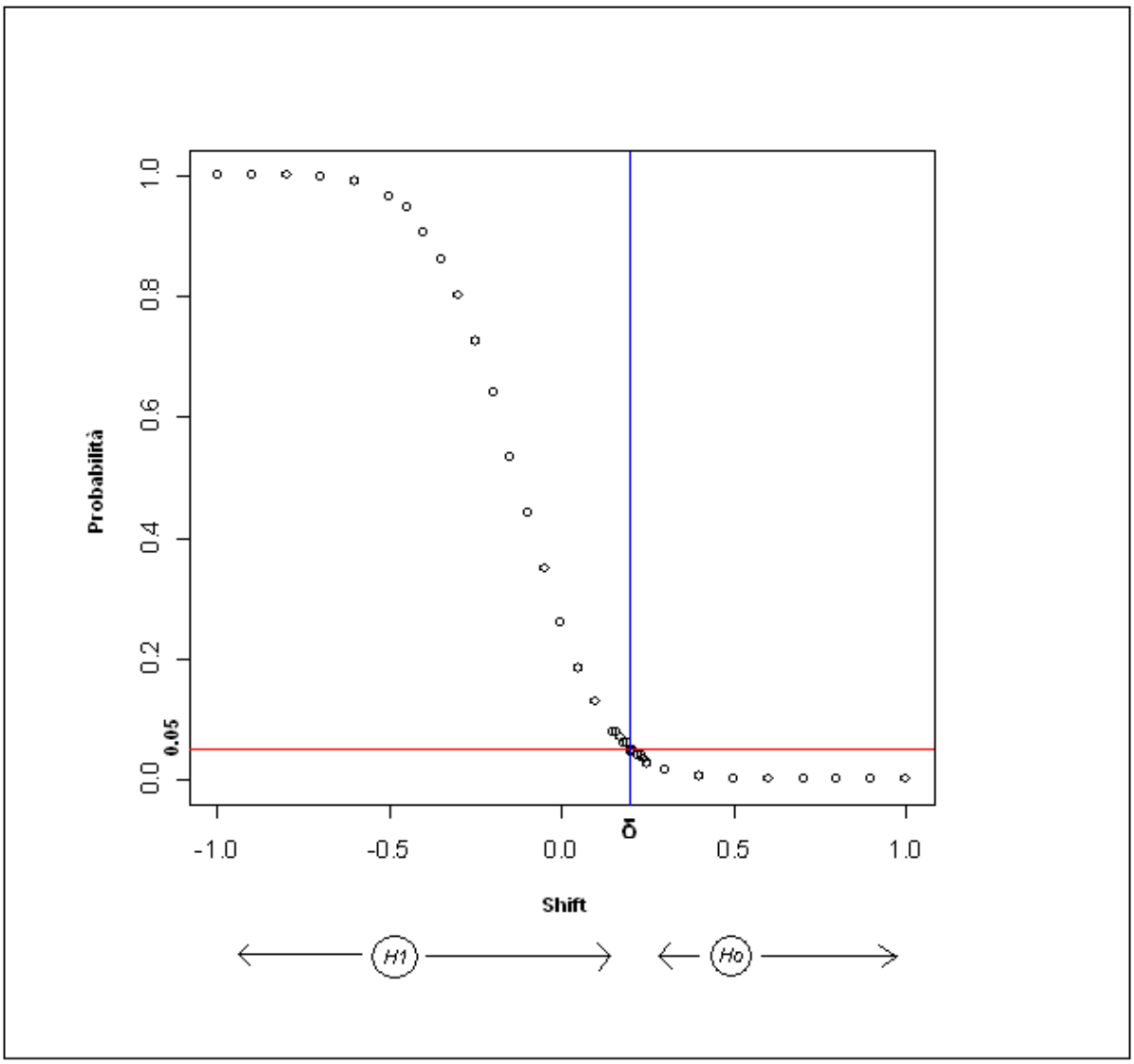


Figura 1: i punti rappresentano le probabilità di accettazione dell'ipotesi alternativa H_1 per diversi valori della vera differenza (shift) delle due medie degli effetti dei due trattamenti, di controllo e sperimentale; le suddette probabilità sono state calcolate come numero totale dei p-value inferiori al valore scelto per α=0.05 sul totale dei 10000 p-value calcolati per le altrettante replicazioni effettuate nello studio.

Nel grafico sono riportate per maggiore comodità di lettura una linea di colore rosso indicante il livello di probabilità $\alpha = 0.05$ ed una linea di colore blu che indica invece il margine di non-inferiorità δ: ancora di più si può notare qui come la probabilità in questione risulti maggiore del livello α sotto H_1, si avvicini allo stesso valore per differenze vicine al margine di non inferiorità, per poi scendere al di sotto della linea rossa nella quando è vera l'ipotesi nulla.

CAPITOLO 2

I TEST DIREZIONALI A DUE CODE

Il tipo di test che verrà presentato e analizzato in questo capitolo rappresenta una soluzione molto interessante al problema del confronto tra popolazioni in ambito clinico. Come già anticipato in sede di introduzione al presente lavoro, esistono diverse tipologie di approccio alla questione che ci poniamo qui: l'inferenza classica propone da una parte il test non-direzionale a due code, che contrappone all'ipotesi nulla di uguaglianza tra gli effetti dei due trattamenti quella alternativa di disuguaglianza tra di essi, e dall'altra parte il test direzionale ad una coda, di cui il test di non-inferiorità analizzato nel primo capitolo rappresenta un caso particolare. Esiste poi un'alternativa, che vedremo essere assolutamente valida, a questi due test che deriva dalla necessità, in un campo così delicato come quello della medicina e delle sperimentazioni cliniche per la ricerca sui trattamenti farmacologici, di tenere in considerazione tutti gli errori che è possibile commettere quanto si conduce una verifica di ipotesi e si prende una decisione scartandone un'altra; come sappiamo la statistica classica prende in considerazione due tipi di errori che si possono verificare facendo inferenza tramite uno dei due test accennati poco sopra: l'errore α del I tipo che consiste nel rifiutare l'ipotesi nulla in favore di quella alternativa quando invece è vera la prima, e l'errore β del II tipo che si commette quando si accetta l'ipotesi nulla quando in realtà essa è falsa. Soffermandosi ad analizzare in particolare il test direzionale ad una coda è facile vedere come questi due tipi di errore non comprendano una terza possibile decisione sbagliata data dal caso in cui si sceglie l'ipotesi alternativa (per esempio data da $\mu_1 > \mu_2$) rifiutando l'ipotesi nulla di uguaglianza dei due trattamenti quando invece è vera una terza possibile ipotesi, e cioè che $\mu_1 < \mu_2$: è proprio da qui che in molti campi di applicazione nasce la necessità di utilizzare un terzo tipo di test, il test direzionale a due code, che sia in grado di distinguere la differenza tra due trattamenti (in entrambe le direzioni possibili) dalla situazione di uguaglianza e che prenda in considerazione un terzo tipo di errore che

fornisca l’entità di rischio di affermare vera una direzione quando invece è giusta quella opposta.

Il presente capitolo sarà dedicato a questo tipo di test: nel primo paragrafo verrà chiarito il concetto di errore del III tipo attraverso i confronto tra i tre tipi di test sopra accennati, verranno poi nel secondo paragrafo formalizzate le ipotesi del test in questione e discusse le caratteristiche principali, la questione legata alla scelta tra l’uso del test o dell’intervallo di confidenza e verranno in ultimo presentati i risultati relativi ad uno studio di simulazione condotto sul test in questione. Infine il terzo paragrafo sarà dedicato alla discussione di alcune questioni legate all’utilizzo di questo tipo di test, anche in relazione ai test classici sopra citati.

2.1. L'ERRORE DEL TERZO TIPO

2.1.1. Tre test a confronto

Per introdurre al meglio l'argomento e cogliere nel modo corretto il vantaggio effettivamente apportato dall'utilizzo del test direzionale a due code credo sia necessario fare una presentazione-confronto dei tre test citati in sede di introduzione a questo capitolo: il test non direzionale a due code, e i test direzionali a una e due code, di cui si formalizzeranno le ipotesi nulla e alternativa, discutendone il significato e il valore in campo clinico. Verranno presentati qui i tre test in modo sintetico, in quanto un paragrafo a parte verrà dedicato alla formalizzazione del test oggetto del capitolo.

Verranno utilizzate per dare continuità al lavoro le stesse notazioni presenti nel primo capitolo per indicare le variabili indicatrici degli effetti dei trattamenti sperimentale e di controllo.

- IL TEST NON DIREZIONALE A DUE CODE:

  questo test viene utilizzato per verificare le seguenti ipotesi

  $$\begin{cases} H_0: & T = C \\ H_1: & T \neq C \end{cases}$$

  dove quindi l'ipotesi nulla di uguaglianza dei due trattamenti viene contrapposta ad un'ipotesi alternativa bilaterale. La regione di rifiuto è qui data da valori estremi del test sia sulla coda destra che su quella sinistra della sua distribuzione.

  Appare chiaro, come era già stato notato in precedenza durante la trattazione del lavoro, come questo tipo di test sia molto poco utile in sede di esperimenti clinici, in quanto uno degli scopi principali degli studi fatti per confrontare diversi trattamenti è sicuramente quello di stabilire il segno della differenza tra i risultati dei due farmaci considerati. Per fare ciò quindi diventa necessaria, una volta applicato questo test e nel caso di rifiuto dell'ipotesi nulla, l'ulteriore applicazione di un opportuno test direzionale per determinare la superiorità del trattamento sperimentale su quello di controllo

o viceversa, con conseguente perdita di informatività dello studio, soprattutto per quanto riguarda il controllo degli errori inferenziali.

- IL TEST DIREZIONALE AD UNA CODA:

  le ipotesi da testare sono le seguenti

$$\begin{cases} H_2: & C \geq T \\ H_3: & C < T \end{cases}$$

  dove l'ipotesi alternativa rappresenta la superiorità del trattamento sperimentale su quello di controllo, mentre l'ipotesi nulla contiene le altre due possibili alternative, e cioè che il nuovo trattamento risulti peggiore o uguale allo standard.

  In questo caso la regione di rifiuto non è più divisa nelle due code della distribuzione del test, ma giace interamente sulla coda sinistra, e rifiutando l'ipotesi nulla per quella alternativa allora noi prendiamo una decisione sul segno della differenza.

- IL TEST DIREZIONALE A DUE CODE:

  le ipotesi che caratterizzano questo test sono le seguenti[19]

$$\begin{cases} H_4: & C > T \\ H_0: & C = T \\ H_3: & C < T \end{cases}$$

  Non si tratta più allora di decidere tra due possibilità, qui le ipotesi sono tre, l'ipotesi nulla H_0 rimane il caso di uguaglianza tra i due trattamenti, mentre l'ipotesi alternativa H_1 del test non direzionale a due code si scinde in due sotto ipotesi direzionali, che indicano una la superiorità del trattamento di controllo rispetto a quello sperimentale e l'altra viceversa la superiorità del nuovo farmaco rispetto a quello preso come standard. Anche in questo caso allora vi saranno due regioni di rifiuto sulle estremità delle code della distribuzione del test, ma qui esse non avranno lo stesso significato portando entrambe a rifiutare l'ipotesi nulla per un'unica ipotesi alternativa,

[19] Kaiser H. F.; "Directional statistical decisions", in *Psychological Review*, (pp 161-162-163-164); 1960.

bensì bisognerà tenere in considerazione in quale delle due cadrà il valore del test per decidere quale delle due ipotesi alternative considerare vera.
La figura 2 riporta le regioni di accettazione e di rifiuto per i tre test, in particolare considerando il test *t* di Student, come indicato in legenda[20]:

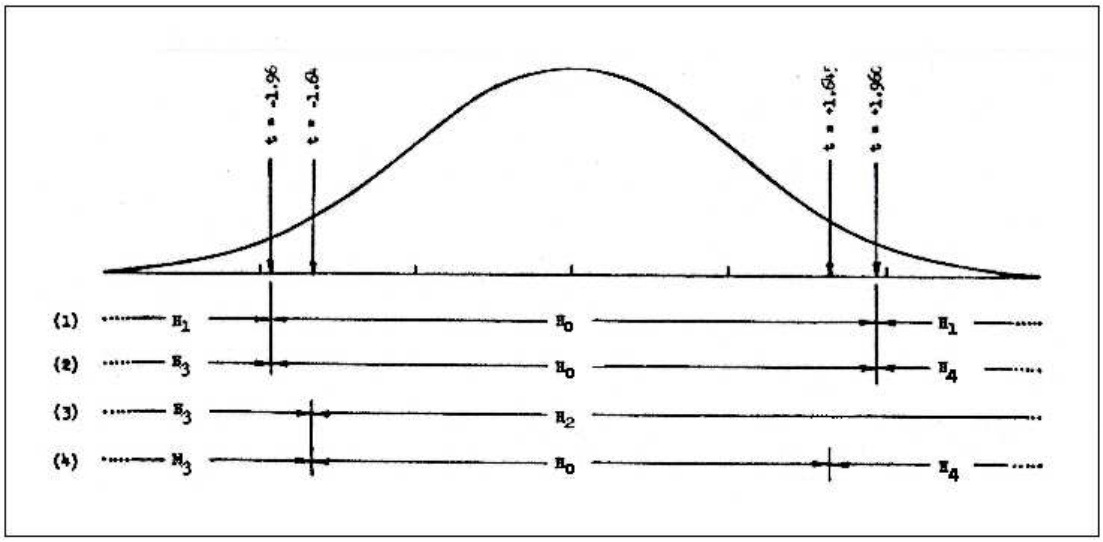


Figura 2: valori critici e regioni di accettazione e di rifiuto per i test direzionale ad una e due code e non direzionale a due code; i numeri sottostanti la figura indicano a quale test sono relativi i risultati riportati, come i seguito descritto:

1. il test non direzionale a due code con $\alpha = 0.05$;
2. il test direzionale a due code con $\alpha_1 = \alpha_2 = 0.025$;
3. il test direzionale ad una coda con $\alpha = 0.05$;
4. il test direzionale a due code con $\alpha_1 = \alpha_2 = 0.05$.

2.1.2. Tre errori associati a tre possibili decisioni sbagliate

Come sappiamo l'inferenza classica tiene in considerazione due tipi di errori legati al fatto di prendere la decisione sbagliata accettando un'ipotesi quando invece è vera l'altra: l'errore α del I tipo e l'errore β del II tipo che sono già stati descritti in fase di presentazione del presente capitolo. L'introduzione del test direzionale a due code come definito qui sopra introduce automaticamente,

[20] Kaiser H. F.; "Directional statistical decisions", in *Psychological Review*, (p 166/Fig. 5); 1960.

come è facile intuire, la presenza di un terzo tipo di errore che consideri la possibilità di rifiutare giustamente l'ipotesi nulla scegliendo però l'ipotesi alternativa che rappresenta la direzione sbagliata. In altre parole questo tipo di errore consiste in un errore di direzione. I primi due test prevedono invece solamente i primi due tipi di errori.
Le quattro possibili situazioni che si possono verificare durante un test come i primi due analizzati sono rappresentate nella tabella 4:

		Stato di natura	
		H_0	H_1
Decisione presa	H_0	decisione corretta	errore β
	H_1	errore α	decisione corretta

Tabella 4: possibilità di decisione durante una verifica di ipotesi per test direzionali ad una coda e test non direzionali a due code.

Le due colonne descrivono lo stato di natura realmente esistente, mentre le righe indicano la decisione presa con la verifica di ipotesi: vi sono due casi in cui si prende la decisione corretta e due in cui invece si commette un errore; di conseguenza si avrà che la potenza del test sarà data da $1-\beta$, probabilità con la quale si rifiuta l'ipotesi nulla quando effettivamente essa è falsa, e $1-\alpha$ sarà la probabilità di accettare l'ipotesi nulla quando essa effettivamente è vera.
Per il test direzionale a due code invece i possibili errori sono sei, spiegati in modo schematico nella tabella 5:

		Stato di natura		
		H_4	H_0	H_3
Decisione presa	H_4	decisione corretta	errore α(1)	errore γ(2)
	H_0	errore β(1)	decisione corretta	errore β(2)
	H_3	errore γ(1)	errore α(2)	decisione corretta

Tabella 5: possibilità di decisione durante una verifica di ipotesi per il test direzionale a due code.

In questo caso l'errore α del I tipo è diviso in due errori in quanto nel caso in cui l'ipotesi nulla sia vera si può rifiutarla sia in favore della prima ipotesi alternativa (errore α(1)) che della seconda (errore α(2)); lo stesso vale per l'errore β del II tipo, che si divide nei due casi in cui l'ipotesi nulla viene accettata anche se falsa perché è invece vera la prima ipotesi alternativa (errore β(1)) o la seconda (errore β(2)). È presente inoltre un terzo tipo di errore, che come vediamo consiste nell'accettare un'ipotesi alternativa quando invece è vera l'altra (errori γ(1) e γ(2)): questi errori sono detti errori del terzo tipo, e sono tenuti i considerazione solamente da questo tipo di test.

2.1.3. Performance a confronto

La differenza esistente tra i test sopra descritti non è evidentemente solamente dovuta alla diversa formulazione delle ipotesi, ma, appare logico, la presenza di una terza possibilità di errore produce un effetto sulla potenza del test direzionale a due code rispetto alla sua versione non direzionale.

Nella figura 3 sono riportate le due curve che descrivono ciascuna la probabilità di scegliere le due possibili ipotesi del test non direzionale a due code a seconda dello stato di natura in cui ci si trova[21]; si noti che le due curve sono complementari:

[21] Kaiser H. F.; "Directional statistical decisions", in *Psychological Review*, (p 163/Fig. 1); 1960.

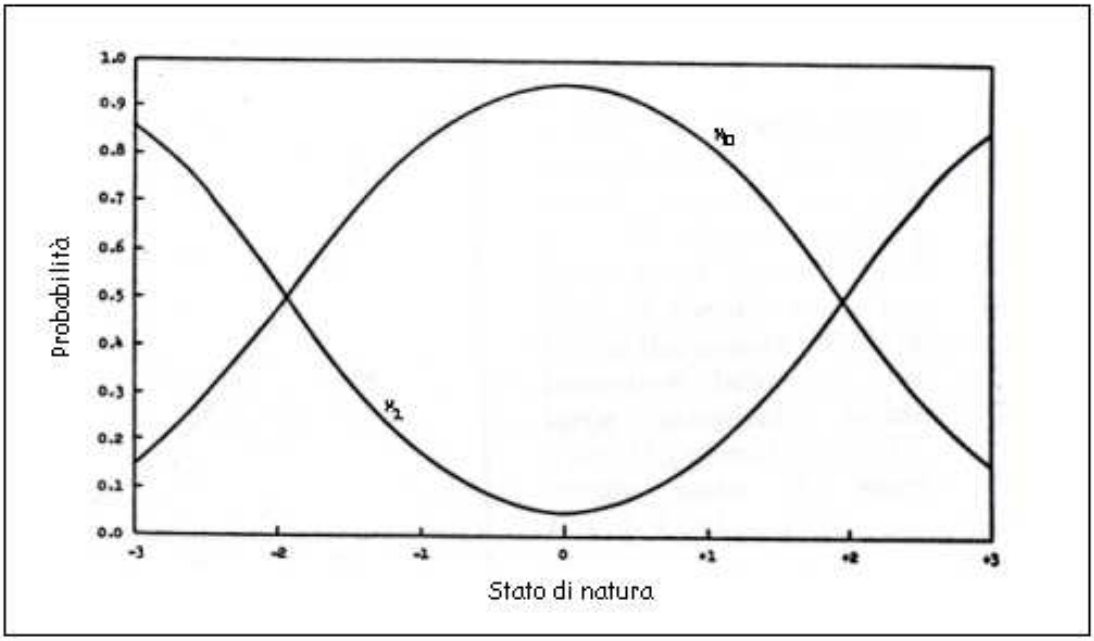


Figura 3: probabilità di scegliere le due possibili ipotesi del test non direzionale a due code a seconda dello stato di natura in cui ci si trova: è stato considerato il test *t*, ad un livello di significatività pari al 5%, e lo stato di natura è stato espresso in unità dell'errore standard della differenza tra le medie dei due trattamenti.

Nella figura 4 sono riportare le analoghe curve (questa volta ovviamente sono tre) per quanto riguarda il test direzionale a due code[22],:

[22] Kaiser H. F.; "Directional statistical decisions", in *Psychological Review*, (p 164/Fig. 2); 1960.

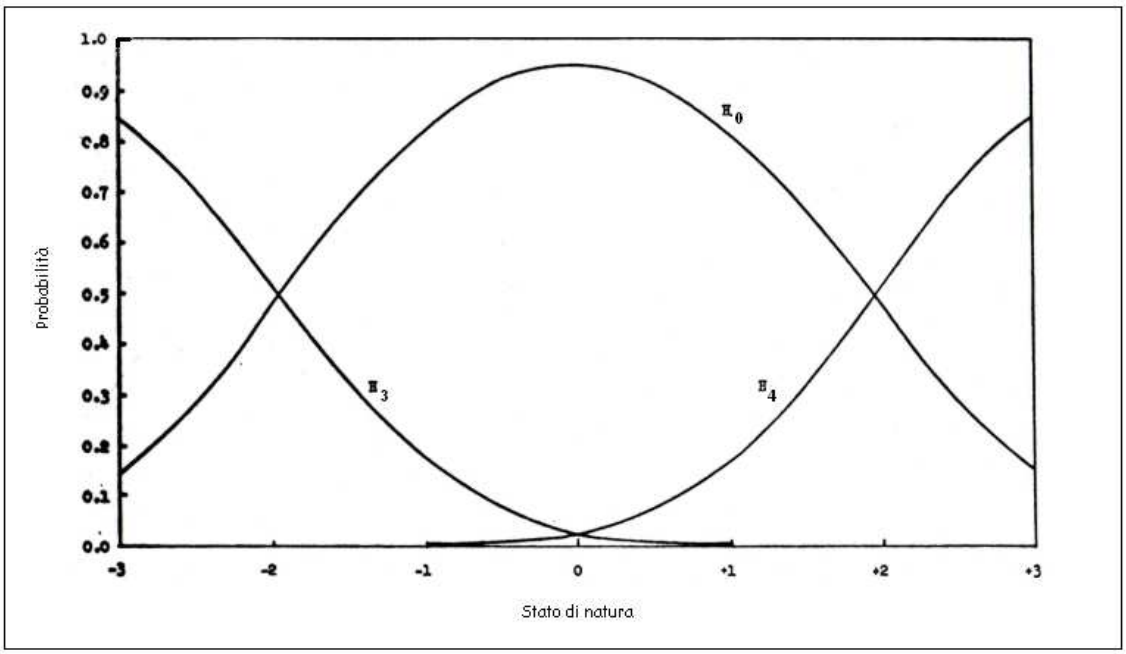


Figura 4: probabilità di scegliere le tre possibili ipotesi del test direzionale a due code a seconda dello stato di natura in cui ci si trova: è stato considerato il test *t*, ad un livello di significatività pari al 5%, e lo stato di natura è stato espresso in unità dello standard error della differenza tra le medie dei due trattamenti; il test è stato costruito ponendo gli errori α(1) e α(2) uguali e pari a 0.025.

Come possiamo vedere mentre la curva relativa alla scelta dell'ipotesi nulla H_0 rimane invariata, vediamo che ciascuna delle due curve relative alle due ipotesi alternative risulta essere minore della curva relativa all'ipotesi alternativa H_1 in figura 3. Come vedremo questa perdita di potenza del test è esattamente pari alla probabilità di commettere un errore del III tipo.

Per quanto riguarda un confronto con il test di superiorità, guardando la figura 5 possiamo notare come questo sia più potente del test direzionale a due code per lo stesso livello di significatività α= α(1)+α(2) (con α(1)+α(2) uguali e pari ad α/2)[23]:

[23] Kaiser H. F.; "Directional statistical decisions", in *Psychological Review*, (p 165/Fig. 3); 1960.

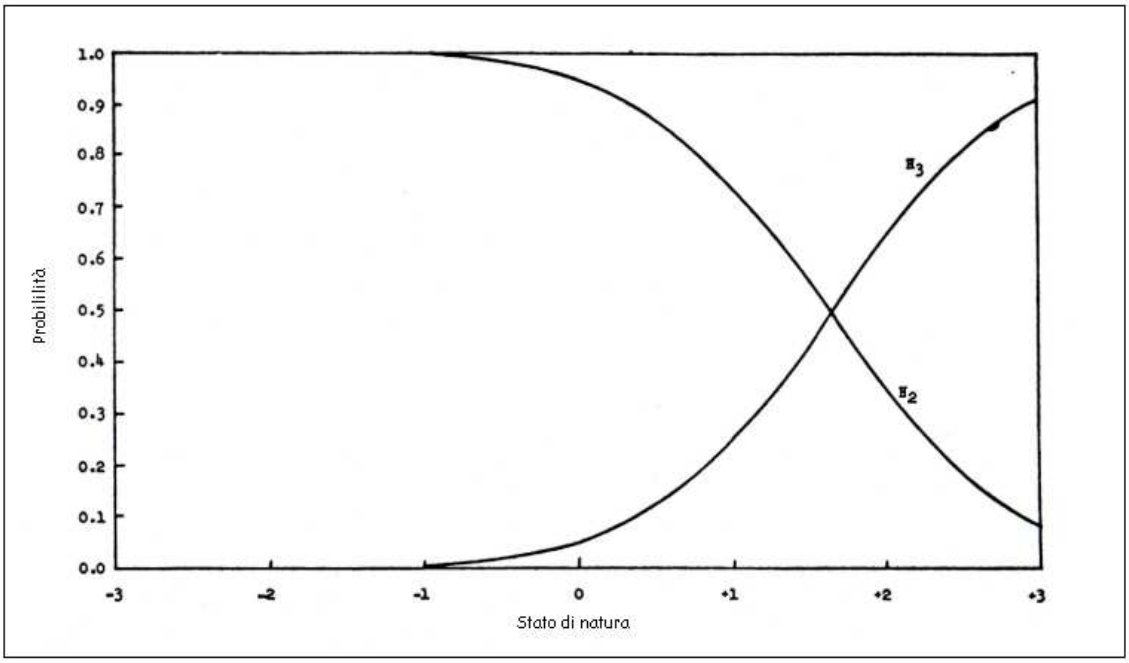


Figura 5: probabilità di scegliere le due possibili ipotesi del test direzionale ad una coda a seconda dello stato di natura in cui ci si trova: è stato considerato il test *t*, ad un livello di significatività pari al 5%, e lo stato di natura è stato espresso in unità dello standard error della differenza tra le medie dei due trattamenti.

La figura 6 invece mostra come per eguagliare la potenza dei due test sia sufficiente porre $\alpha = \alpha(1)=\alpha(2)$[24]; questo però va a raddoppiare la probabilità di commettere un errore del I tipo per il test direzionale a due code, riducendo così l'affidabilità del test stesso:

[24] Kaiser H. F.; "Directional statistical decisions", in *Psychological Review*, (p 165/Fig. 4); 1960.

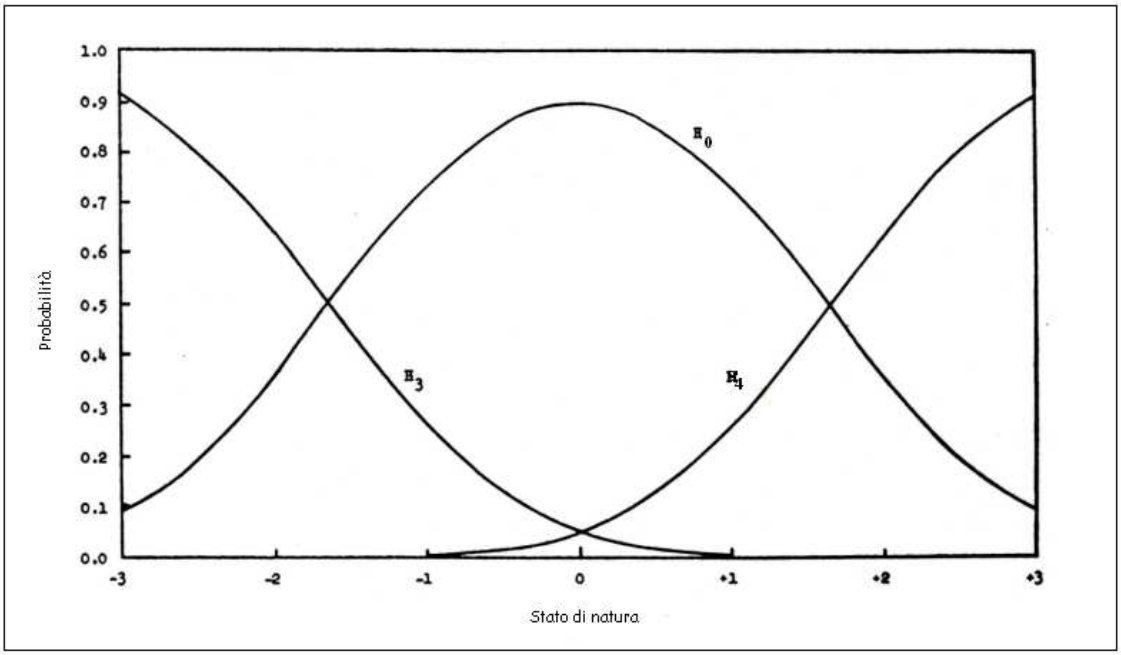


Figura 6: probabilità di scegliere le tre possibili ipotesi del test direzionale a due code a seconda dello stato di natura in cui ci si trova: è stato considerato il test *t*, ad un livello di significatività pari al 5%, e lo stato di natura è stato espresso in unità dello standard error della differenza tra le medie dei due trattamenti; il test è stato costruito ponendo gli errori α(1) e α(2) uguali e pari a 0.05.

Dati relativi alle probabilità di accettazione delle ipotesi alternative del test direzionale a due code saranno riportati anche nel paragrafo dedicato allo studio di simulazione condotto sul test in questione.

Comunque nonostante questa perdita di potenza del test direzionale a due code rispetto agli altri due considerati, rimane comunque innegabile la convenienza e la maggiore desiderabilità di questo tipo di approccio rispetto agli altri due presi in considerazione, in quanto esso è l'unico, come vedremo più nello specifico, in grado di considerare tutti i possibili stati di natura e, a differenza del test di superiorità, di decidere per una delle ipotesi alternative non per esclusione dell'altra, ma per accettazione della stessa.

2.2. IL TEST

2.2.1. Due possibili approcci

Credo sia necessario riprendere in mano le motivazioni per l'utilizzo del test direzionale a due code: come abbiamo già accennato in ambito di confronto tra trattamenti per valutare la differenza di efficacia, l'inferenza statistica classica propone principalmente due test: il test non-direzionale a due code e il test direzionale ad una coda, descritti all'inizio del paragrafo precedente, che però sembrano essere limitati per la risoluzione di alcune problematiche tipiche degli studi clinici in questione. In particolare, come già è stato fatto notare in precedenza il test non direzionale a due code non è in grado, per sua stessa costruzione , di prendere decisioni sul segno della differenza di efficacia dei due trattamenti analizzati, ma riesce solamente a confermare o escludere tale differenza; il test cosiddetto di superiorità dall'altra parte nel caso di rifiuto dell'ipotesi nulla in favore di quella alternativa da un giudizio di tipo direzionale sulla differenza tra i due trattamenti, ma non riesce a tenere sotto controllo tutta la parte opposta della distribuzione del test stesso, non essendo quindi in grado di dire nulla in merito ad eventuali differenze di segno opposto a quello ipotizzato al momento della formalizzazione delle ipotesi da testare.

La soluzione ideale a questa problematica sarebbe quella di effettuare prima un test non direzionale a due code per testare l'esistenza o meno di una differenza statisticamente significativa, e nel caso di rifiuto dell'ipotesi nulla applicare un ulteriore test, questa volta di superiorità, impostando le ipotesi di conseguenza a quanto ottenuto con il primo test, per testare con maggior potenza l'eventuale superiorità di un trattamento rispetto all'altro. Questo tipo di soluzione, anche se spesso adottata da diversi testi statistici, porta con sé alcune problematiche da non sottovalutare quali la differenza di potenza e quindi di ampiezza campionaria richiesta per i due diversi test, e soprattutto la mancata considerazione di quello che nel paragrafo precedente abbiamo definito come errore del III tipo.

Tutte queste premesse hanno portato alla formulazione del test direzionale a due code, in grado di risolvere le questioni fin qui sollevate. Per meglio

comprendere il suddetto test mi è sembrato opportuno riportare due diversi punti di vista dello stesso test, il primo che riprende il concetto di test di superiorità e vede il test direzionale a due code come un duplice test di superiorità che testa prima una direzione e poi la direzione opposta, e il secondo che vede il test composto da tre ipotesi, come è stato presentato nel paragrafo precedente.

La prima formulazione del test in questione è costituita dalle seguenti ipotesi, che costituiscono di fatto le ipotesi di due test di superiorità condotti simultaneamente[25]:

$$\begin{cases} H_2: & C \leq T \\ H_3: & C > T \end{cases}$$

e

$$\begin{cases} H_2: & C \geq T \\ H_3: & C < T \end{cases}$$

I due test di superiorità qui sopra formalizzati devono avere livelli di significatività la cui somma sia pari ad α perchè il test che ne risulti abbia tale livello di significatività, anche se non è necessario che i singoli livelli di significatività siano uguali, ed essi possono essere decisi diversi a seconda dello studio che si sta conducendo e degli obiettivi che ci si pone con esso. Il livello α che ne deriva allora è definito come la probabilità di rigettare almeno una delle due ipotesi nulle quando entrambe sono invece vere.

Dal secondo punto di vista si può vedere il test direzionale a due code come un unico test, cioè, riportando la formulazione introdotta nel paragrafo precedente:

$$\begin{cases} H_4: & C > T \\ H_0: & C = T \\ H_3: & C < T \end{cases}$$

In seguito, per rimanere in linea con il paragrafo precedente, adotteremo questa formulazione per il test in questione.

Come già anticipato nel paragrafo precedente, e mostrato in figura 2, il test in questione ha due valori critici che per lo stesso livello di significatività equivalgono ai valori critici di un normale test non direzionale a due code. Nelle

[25] Leventhal L., Huynh C. L.; "Directional decisions for two-tailed tests: Power, error rates and sample size", in *Psychological Methods*, (p 279); 1996.

prossime sezioni del paragrafo verranno descritte tutte le altre caratteristiche per cui i due test differiscono e che rendono particolare il test direzionale a due code, quali la potenza, gli errori, l'ampiezza campionaria.

2.2.2. I possibili errori

Per costruzione il test direzionale a due code prevede la presenza di tre stati di natura esaustivi e mutuamente esclusivi (H_0, H_4, H_3) e di tre decisioni anch'esse esaustive e mutuamente esclusive (accettare H_0, accettare H_4, accettare H_3).

Possiamo riportare gli errori relativi alle varie decisioni rispettivamente ai diversi stati di natura che si possono verificare attraverso una versione più precisa della tabella mostrata nel paragrafo precedente:

		Stato di natura		
		H4	H0	H3
Decisione presa	H4	decisione corretta 1-β(1)-γ(1) *potenza(4-4)*	errore del I tipo α(1)	errore del III tipo γ(2) potenza(4-3)
	H0	errore del II tipo β(1)	decisione corretta 1-α(1)-α(2)	errore del II tipo β(2)
	H3	errore del III tipo γ(1) *potenza(3-4)*	errore del I tipo α(2)	decisione corretta 1-β(2)-γ(2) *potenza(3-3)*

Tabella 6: possibilità di decisione durante una verifica di ipotesi per il test direzionale a due code; si noti che, in modo concettualmente errato, sono stati indicati come potenza del test anche gli errori del terzo tipo.

Nella tabella 6 abbiamo indicati i tre tipi di errori:

i. l'errore del I tipo (α(1) e α(2)) che si commette rifiutando l'ipotesi nulla quando invece essa è vera;
ii. l'errore del II tipo (β(1) e β(2)) che si commette quando si accetta l'ipotesi nulla quando invece è vera una delle due ipotesi alternative;
iii. l'errore del III tipo (γ(1) e γ(2)) che si commette quando si accetta una delle due ipotesi alternative quando invece è vera l'altra.

Appare chiaro dalla tabella come gli errori del III tipo non possano essere considerati, per costruzione degli stessi, né dal test non direzionale a due code né dal test di superiorità, in quanto questi non prevedono proprio il caso in cui si accetti un'ipotesi alternativa quando invece è vera l'altra, non contemplando per niente l'esistenza di una seconda ipotesi alternativa.

2.2.3. La potenza del test

Come abbiamo già anticipato nel paragrafo precedente, si ha con l'utilizzo di questo tipo di test una perdita di potenza rispetto agli altri due test presi come confronto in questo capitolo.

La potenza di un test è solitamente definita come la probabilità di rifiutare l'ipotesi nulla in favore di quella alternativa quando la prima è falsa; seguendo alla lettera questa definizione di potenza e guardando il test direzionale a due code (per comodità riferiamoci alla tabella 6) ci accorgiamo che ci sono quattro diverse situazioni in cui accade che l'ipotesi nulla, da falsa, viene rifiutata, in quanto può accadere che:

- si rifiuti H_0 in favore di H_4 quando è vera quest'ultima;
- si rifiuti H_0 in favore di H_4 quando invece è vera H_3;
- si rifiuti H_0 in favore di H_3 quando è vera quest'ultima;
- si rifiuti H_0 in favore di H_3 quando invece è vera H_4;

Il primo e il terzo caso rientrano nella definizione standard di potenza di un test, e sono casi che si possono riscontrare anche in un classico test a due ipotesi (nulla ed alternativa); il secondo e il quarto caso invece sono tipi di situazioni riscontrabili solo nel test in questione, e dovute alla struttura stessa del test: rifiutare un'ipotesi nulla falsa in favore dell'ipotesi alternativa sbagliata

rappresenta, come sappiamo l'errore del III tipo, esclusivo del test direzionale a due code qui analizzato. Il rifiuto dell'ipotesi nulla falsa però fa rientrare questo tipo di decisione nella definizione di potenza del test (come mostrano le celle in alto a destra e in basso a sinistra della tabella 6: questo ci porta alla necessità di distinguere la definizione di potenza nel caso di test standard (inteso come test avente un'ipotesi nulla ed un'unica ipotesi alternativa) e nel caso particolare del test direzionale a due code, distinzione inevitabile in quanto altrimenti nel test a tre ipotesi si finirebbe con l'includere nella potenza del test anche decisioni non corrette.

Inoltre la definizione tradizionale di potenza del test applicata al test direzionale a due code produrrebbe un valore della potenza del test stesso pari alla potenza di un test non direzionale a due code con lo stesso livello di significatività, in quanto i valori critici per l'accettazione o il rifiuto dell'ipotesi nulla rimangono gli stessi.

La definizione alternativa[26] da utilizzare vede invece la potenza di un test come la probabilità (condizionata) di rifiutare una falsa ipotesi nulla in favore della giusta ipotesi alternativa, aggiungendo così la condizione di accettare la giusta ipotesi alternativa, cosa che non avveniva nella definizione classica menzionata prima (anche perché inutile in quanto il rifiuto dell'ipotesi nulla nei test a due ipotesi equivale automaticamente all'accettazione dell'unica ipotesi alternativa esistente).

Con questa nuova definizione di potenza vengono esclusi dalla potenza del test stesso gli errori del III tipo, dando luogo alla seguente tabella:

[26] Leventhal L., Huynh C. L.; "Directional decisions for two-tailed tests: Power, error rates and sample size", in *Psychological Methods*, (p 282); 1996.

		Stato di natura		
		H4	H0	H3
Decisione presa	H4	decisione corretta 1-β(1)-γ(1) *potenza(4-4)*	errore del I tipo α(1)	errore del III tipo γ(2)
	H0	errore del II tipo β(1)	decisione corretta 1-α(1)-α(2)	errore del II tipo β(2)
	H3	errore del III tipo γ(1)	errore del I tipo α(2)	decisione corretta 1-β(2)-γ(2) *potenza(3-3)*

Tabella 7: possibilità di decisione durante una verifica di ipotesi per il test direzionale a due code; si noti che, in modo concettualmente corretto, questa volta gli errori del terzo tipo non sono stati indicati come potenza del test.

Calcolando la potenza del test direzionale a due code tramite questa seconda definizione, intuitivamente più corretta della prima nel nostro particolare caso, si conferma il risultato visto nel paragrafo 2.1.3.: il test risulta essere meno potente del corrispondente test non direzionale a due code. Con la definizione tradizionale si avrebbe che la potenza sotto H_4 è data da:

$$potenza(H_4) = potenza(4-4) + potenza(4-3) = 1-\beta(1)-\gamma(1)+\gamma(1) = 1-\beta(1)$$

e la potenza sotto H_3 è data da:

$$potenza(H_3) = potenza(3-3) + potenza(3-4) = 1-\beta(2)-\gamma(2)+\gamma(2) = 1-\beta(2)$$

e quindi la potenza del test risulterebbe pari a:

$$potenza = 1-\beta$$

per un dato stato di natura.

Con la definizione alternativa invece si ha che la potenza del test direzionale a due code è data semplicemente da, sotto H_4:

$$potenza(H_4) = potenza(4-4) = 1-\beta(1)-\gamma(1)$$

e sotto H_3 da:

$$potenza(H_3) = potenza(3-3) = 1-\beta(2)-\gamma(2)$$

in quanto nei due casi rispettivamente $potenza(4-3)$ e $potenza(3-4)$ non vengono considerate come potenza da conteggiare, e quindi la potenza del test effettiva risulta pari a:

$$potenza = 1-\beta-\gamma$$

per un dato stato di natura.

Quindi in generale possiamo affermare, come era stato già accennato in precedenza, che la "perdita di potenza" del test a tre ipotesi rispetto al test non direzionale a due code è esattamente pari all'ammontare dell'errore del III tipo. Per quanto riguarda la differenza di potenza tra il test direzionale a due code ed il test direzionale ad una coda, differenza che vede naturalmente in vantaggio quest'ultimo test, possiamo affermare che per ottenere la stessa potenza di un test di superiorità ad un livello di significatività pari ad $\alpha/2$ è necessario condurre il test a tre scelte ad un livello di significatività doppio pari ad α, e quindi con una grande perdita di affidabilità.

Tuttavia non è del tutto corretto parlare di "perdita di potenza" del test a tre code rispetto agli altri due(ho preferito infatti scriverlo tra due virgolette), in quanto le due procedure operano in condizioni e con obiettivi inferenziali totalmente diversi tra loro. Inoltre abbiamo visto come sia stato necessario ricorrere ad una definizione particolare, per quanto riguarda la potenza del test direzionale a due code, rispetto alla definizione classica di potenza che viene considerata quando si parla di test direzionale ad una coda o di test non direzionale a due code; risulta quindi poco verosimile un confronto diretto tra le potenze dei test in questione.

In figura 7 sono riportati alcuni esempi grafici che riassumono quanto detto fin qui relativamente alla potenza dei test direzionale e non direzionale a due code, e anche del test direzionale ad una coda, come descritto in legenda[27]:

[27] Leventhal L., Huynh C. L.; "Directional decisions for two-tailed tests: Power, error rates and sample size", in *Psychological Methods*, (p 283/Fig. 4); 1996.

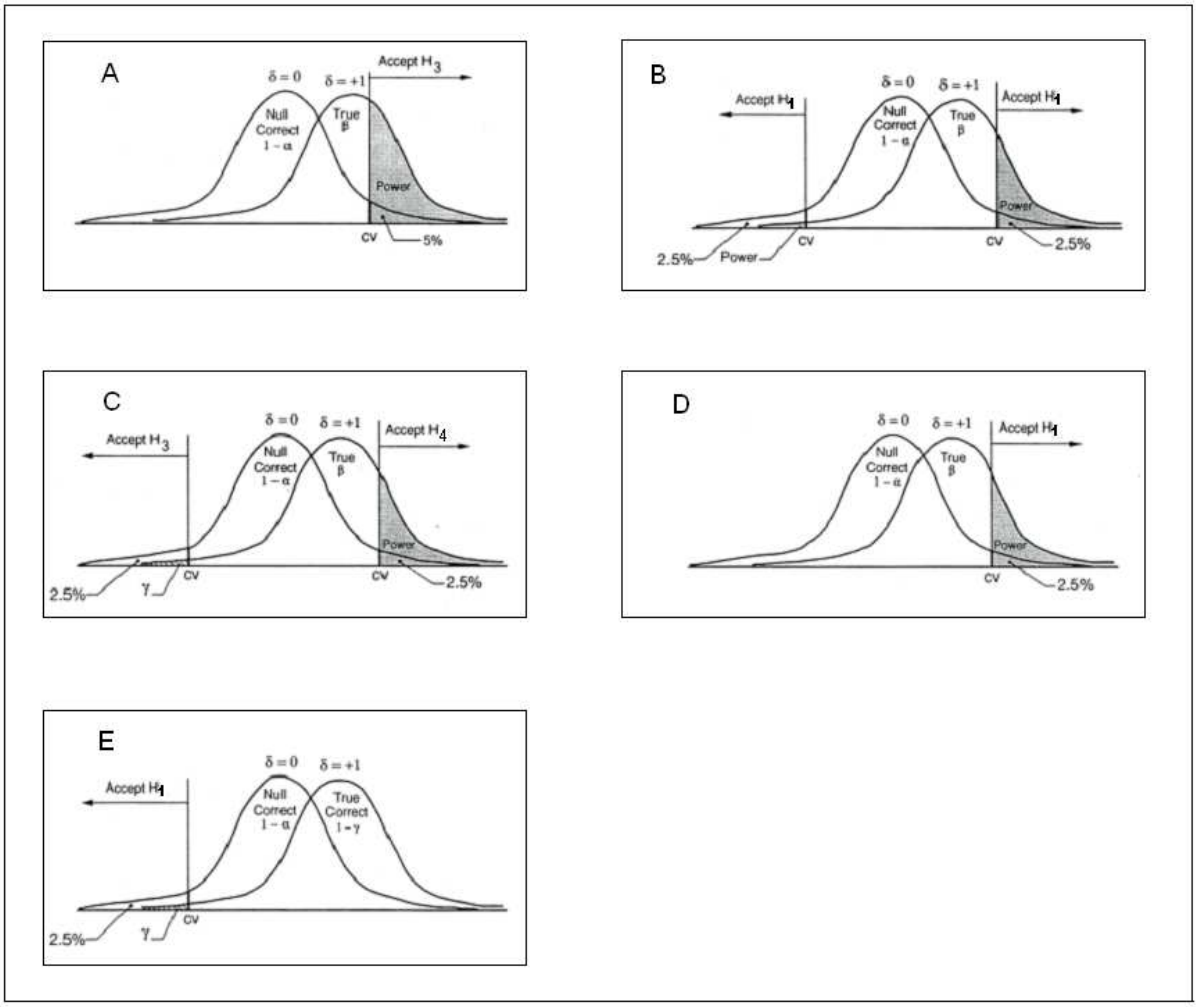


Figura 7: potenza ed errori del I, II e III tipo per i test direzionale ad una e due code e non direzionale a due code; le cinque figure indicano:

A test direzionale ad una coda con $\alpha = 0.05$;

B test non direzionale a due code con $\alpha = 0.05$;

C test direzionale a due code con $\alpha_1 = \alpha_2 = 0.025$;

D test direzionale ad una coda con $\alpha = 0.025$, dove H_1 descrive la direzione corretta;

E test direzionale ad una coda con $\alpha = 0.025$, dove H_1 descrive la direzione sbagliata;

2.2.4. L'ampiezza campionaria

Le considerazioni da fare a proposito dell'ampiezza campionaria derivano direttamente da quanto detto riguardo alla potenza dei tre test presi in considerazione; in generale si deve distinguere il caso in cui si utilizza la definizione tradizionale di potenza del test o quella alternativa definita nel paragrafo precedente:

- nel caso in cui si usi la definizione tradizionale di potenza del test, abbiamo visto che il test più potente risulta essere il test direzionale ad una coda, seguito dai due test a due code, direzionale e non, che fanno registrare la stessa potenza. Di conseguenza il primo test richiede un'ampiezza campionaria minore rispetto a quella equivalente richiesta degli altri due test;
- nel caso di definizione alternativa di potenza, il test direzionale a due code risulta essere il meno potente dei tre, e quindi richiede un'ampiezza campionaria maggiore rispetto al test non-direzionale a due code e al test di superiorità, che rimane comunque il più potente.

Queste considerazioni, che sembrano ovvie dopo quelle fatte riguardo la potenza dei tre test, devono essere tenute in considerazione, in quanto come sappiamo la decisione della numerosità campionaria dovrebbe essere una scelta fatta prima dell'inizio dell'esperimento clinico, e questo significa che è necessario fare in principio una scelta sulla direzionalità della risposta che si vuole ottenere e sul tipo di definizione di potenza del test che si vuole tenere in considerazione.

Riportiamo di seguito, in tabella 7, alcuni risultati derivanti dal confronto tra i tre test finora presi in considerazione[28]:

[28] Leventhal L., Huynh C. L.; "Directional decisions for two-tailed tests: Power, error rates and sample size", in *Psychological Methods*, (p 285/Table 1); 1996.

	One-tailed	Nondirectional two-tailed	Directional two-tailed
α	.05	.05	.05
Test statistic & value	$z = 4.91869$	$z = 4.91869$	$z = 4.91869$
Critical value(s)	$z_\alpha = -1.645$	$z_{\alpha/2} = \pm 1.96$	$z_{\alpha/2} = \pm 1.96$
Decision	Retain H_2	Accept H_1	Accept H_4
Power & error rates			
$\delta = 1, n_1 = 32,$	π = undefined[a]	$\pi = 0.078$	$\pi = 0.071$
$n_2 = 30$	β = undefined[a]	$\beta = 0.922$	$\beta = 0.922$
	$\gamma = 0.0163$[b]	γ = undefined	$\gamma = 0.007$
Sample size			
for $\pi = .10, \delta = 1$	n = undefined[a]	$n = 54.794 \approx 55$	$n = 58.909 \approx 59$
for $\pi = .30, \delta = 1$	n = undefined[a]	$n = 263.426 \approx 264$	$n = 263.788 \approx 264$
for $\pi = .60, \delta = 1$	n = undefined[a]	$n = 627.017 \approx 628$	$n = 627.039 \approx 628$

Tabella 7: confronto tra potenza ed errori per una vera differenza $\delta = 1$ (quindi contraria alla previsione fatta) e le ampiezze campionarie necessarie per ottenere invece una potenza pari a 0.10, 0.30 e infine 0.60, per i tre test finora presi in considerazione, test direzionali ad una e due code e test non direzionale a due code; è stato considerato il caso del test *z*, per la sua semplicità di calcolo e di interpretazione, per il confronto tra le medie di due campioni indipendenti, posto che:

- la vera differenza tra le due medie $\delta = \mu_1 - \mu_2$ è ipotizzata essere negativa, e cioè essere $\mu_1 < \mu_2$;
- le ampiezze campionarie sono $n_1 = 32$ ed $n_2 = 30$;
- le medie campionarie sono risultate essere $\overline{X}_1 = 40$ e $\overline{X}_2 = 30$;
- le due popolazioni di partenza sono distribuite normalmente con deviazione standard uguale e pari a $\sigma_1 = \sigma_2 = \sigma = 8$;
- il livello di significatività scelto è di $\alpha = 0.05$.

2.2.5. Intervalli di confidenza

In alternativa ai test finora utilizzati si può sfruttare, come sappiamo, la costruzione di intervalli di confidenza; anche con questa metodologia è comunque necessario fare una scelta preliminare all'inizio dell'esperimento sulla direzionalità della risposta che si vuole ottenere, che ci indirizzi sulla scelta

del giusto intervallo di confidenza per arrivare con risultati affidabili all'obiettivo dello studio.

È noto infatti che esistono due tipi di intervalli di confidenza, quelli di tipo bilaterale, che vengono quindi costruiti intorno al valore stimato per la grandezza su cui si sta indagando, e quelli unilaterali, che presentano solamente un estremo inferiore oppure un estremo superiore e sono aperti nelle rispettive direzioni opposte.

È facile intuire come il primo tipo di intervallo possa essere avvicinato da un punto di vista concettuale al test direzionale a due code, in quanto in grado di fornire informazioni sul segno della differenza di efficacia esistente tra i due trattamenti, in entrambe le direzioni. Il secondo tipo di intervallo invece risulta per costruzione adatto allo studio del segno della suddetta differenza in una direzione ipotizzata preventivamente alla costruzione dell'intervallo stesso, e quindi risulta soggetto ad errori del III tipo più pesanti rispetto al primo caso appena descritto.

Il primo intervallo di confidenza è definito, ad un livello dell'(1-α)100%, nel modo seguente:

$$CI = (C - T) \pm z_{1-\alpha/2}\sigma_{(C-T)}$$

e la sua interpretazione è già stata presentata nel corso del primo capitolo: se tutti i valori dell'intervallo stesso risultano essere minori di un certo valore soglia (che può essere 0, oppure genericamente *c* relativamente alle necessità che sorgono durante lo studio che si sta conducendo) allora si può affermare con un livello di confidenza dell'(1-α)100% che la differenza tra i due trattamenti analizzati risulta minore di quella soglia; se viceversa l'intervallo contiene solo valori superiori alla soglia prestabilita, allora si può concludere che la differenza è superore al valore di riferimento. Se infine l'intervallo contiene tra i suoi valori anche quello scelto come soglia allora non si può concludere che la differenza tra i due trattamenti sia diversa dal valore in questione.

Queste regole di interpretazione dell'intervallo bilaterale fanno capire come esso possa svolgere le stesse funzioni di un test direzionale a due code, come si era accennato poco sopra; infatti questa tipologia di intervalli è in grado di prendere decisioni in entrambe le direzioni, tenendo conto quindi dell'errore di

terza specie. Infatti prendiamo in considerazione il test a tre ipotesi, che come sappiamo è definito nel modo seguente:

$$\begin{cases} H_4: & C > T \\ H_0: & C = T \\ H_3: & C < T \end{cases}$$

e che può peraltro essere riscritto, per coerenza con il concetto di differenza che è stato utilizzato per parlare degli intervalli, come:

$$\begin{cases} H_4: & C - T > 0 \\ H_0: & C - T = 0 \\ H_3: & C - T < 0 \end{cases}$$

Allora il caso in cui l'intervallo di confidenza considerato contiene valori tutti inferiori a 0 (in questo caso perché è 0 la soglia considerata), corrisponde all'accettazione dell'ipotesi alternativa H_3; il caso in cui i valori contenuti nell'intervallo siano tutti superiori a 0 a sua volta corrisponde ad accettare l'ipotesi H_4 e infine quando l'intervallo contiene lo 0 ci troviamo ad accettare l'ipotesi nulla H_0.

Gli intervalli di confidenza unilaterali vengono costruiti tenendo solamente un estremo, inferiore o superiore, e lasciandoli aperti nella direzione opposta; si possono quindi avere due tipi di intervallo, limitati superiormente o inferiormente, i cui estremi sono definiti rispettivamente come segue:

$$CI_{\sup} = (C - T) + z_{1-\alpha}\sigma_{(C-T)}$$

$$CI_{\inf} = (C - T) - z_{1-\alpha}\sigma_{(C-T)}$$

Il primo intervallo contiene quindi tutti i valori inferiori di $CI_{\sup}$, mentre quello limitato inferiormente contiene tutti i valori superiori a $CI_{\inf}$; l'interpretazione di questi intervalli è analoga a quella per gli intervalli bilaterali: la presenza di soli valori superiori alla soglia porterà ad affermare la superiorità della differenza dei due trattamenti al valore in riferimento e viceversa la presenza di valori dell'intervallo tutti inferiori alla soglia considerata; se la soglia è compresa nell'intervallo non siamo autorizzati ad affermare che la differenza sia diversa (maggiore o minore rispettivamente) da quel valore. Ci accorgiamo però che la particolare costruzione di questi intervalli richiede che venga ipotizzato un

segno della differenza prima della scelta di quale tra i due utilizzare: un segno positivo della differenza rispetto allo zero considerato potrà essere indagato e confermato (cioè accettata l'ipotesi alternativa di differenza tra i due trattamenti maggiore di zero) solamente attraverso la costruzione di un intervallo limitato inferiormente, in quanto è ovvio che un intervallo limitato superiormente non conterrà il valore soglia solo nel caso in cui il limite superiore risulti inferiore della suddetta soglia e quindi nel caso in cui l'intervallo contenga solamente valori inferiori alla soglia stessa; in questo caso però verrebbe dimostrata, al livello di confidenza considerato, l'inferiorità della differenza tra i trattamenti rispetto alla soglia di riferimento. Il discorso inverso vale nel caso si voglia dimostrare l'inferiorità della differenza tra i due trattamenti alla soglia considerata.

Per questi motivi questa tipologia di intervalli può essere considerata alla stregua del test direzionale ad una coda: anche per questo test infatti è necessario come sappiamo ipotizzare il segno della differenza tra i due trattamenti per formalizzare nel modo corretto le ipotesi, e in relazione al tipo di ipotesi che si è autorizzati a fare si costruiranno due test differenti tra loro, formalizzati nei due modi seguenti:

$$\begin{cases} H_2: & C \le T \\ H_3: & C > T \end{cases}$$

oppure

$$\begin{cases} H_2: & C \ge T \\ H_3: & C < T \end{cases}$$

Di conseguenza, analogamente a quanto detto sul confronto tra test direzionali a una o due code, gli intervalli di confidenza unilaterali sono caratterizzati da un errore del III tipo molto alto se l'ipotesi direzionale utilizzata per la costruzione dell'intervallo è errata, mentre presentano una probabilità più alta di identificare la giusta direzione se impostati nel modo corretto. La scelta tra intervalli unilaterali o bilaterali allora dovrà seguire gli stessi principi già accennati per la scelta tra test direzionale ad una o due code, con la preferenza quindi, nel caso in cui sia obiettivo principale dello studio determinare il segno della differenza tra i due trattamenti, per gli intervalli bilaterali, che consentono di tenere sotto controllo l'errore del III tipo, e quindi di dare risultati in generale più affidabili.

Per quanto riguarda infine la scelta tra l'utilizzo di intervalli di confidenza bilaterali oppure di test a tre ipotesi, entrambi allo stesso livello di confidenza (1-α) portano alla medesima conclusione in merito alla differenza indagata; è anche vero però che l'intervallo di confidenza offre di più del test classico, in quanto esso stima anche la grandezza del parametro di interesse con un certo livello di confidenza e ne da la precisione attraverso l'ampiezza stessa dell'intervallo, cose che non avvengono nel caso dell'uso del test, che di contro è in grado di fornire il *p-value* associato al test stesso (che rimane comunque una cosa ben distinta dall'ordine di grandezza fornito dall'intervallo di confidenza). D'altra parte comunque il test consente di misurare gli errori β e γ.

In conclusione possiamo affermare che conviene dapprima costruire un intervallo di confidenza bilaterale per decidere sul segno della differenza indagata, e conoscere l'ordine di grandezza del parametro stesso, e per completezza di analisi ricorrere in un secondo tempo all'utilizzo di un test direzionale a due code per quanto riguarda *p-value* ed entità degli errori del I e II tipo.

2.2.6. Studio di simulazione

Analogamente a quanto fatto nel primo capitolo con il test di non-inferiorità anche per il test direzionale a due code è stato condotto uno studio di simulazione, con lo scopo di confermare la non distorsione del test in oggetto. È stato utilizzato anche in questo caso il test *t* di Student.

Lo studio è stato fatto ancora su 10000 replicazioni del test su due campioni indipendenti: uno, quello relativo al trattamento di controllo, generato sempre dalla stessa popolazione normale di media e varianza unitarie che era stata utilizzata anche nel primo studio di simulazione presentato nel primo capitolo; per generare il secondo campione sono state considerate questa volta 31 popolazioni di partenza differenti, aventi anche qui varianza comune ed uguale a 1, e medie fatte variare da un minimo di 0 ad un massimo di 2, distanziate da intervalli di una unità verso i due estremi e di 0,5 vicino alla media della popolazione assunta trattata con il farmaco di controllo (cioè vicino a 1) per

osservare meglio il comportamento del test in prossimità del valore critico per lo shift, ossia lo zero.

È bene comunque precisare il modo in cui sono stati calcolati i p-value corrispondenti ai test condotti; sappiamo che il test non direzionale a due code calcola il p-value associato al valore della statistica test ottenuta come segue:

$$p = 2\min\left\langle \Pr\left\{T \le t^{oss}\right\}, \Pr\left\{T \ge t^{oss}\right\}\right\rangle$$

Nel caso del test a tre ipotesi che stiamo qui analizzando sarà evidentemente necessario distinguere tra due diversi p-value: quello associato all'eventuale accettazione della prima ipotesi alternativa "sinistra", cioè H_3: $C < T$, e quello associato all'eventuale accettazione dell'ipotesi alternativa la cui regione di accettazione si trova sulla coda destra, cioè H_4: $C > T$; com'è facilmente intuibile i due p-value sono complementari, il più piccolo è esattamente la metà del p-value calcolato per il test non direzionale a due code, in quanto esso è definito proprio come:

$$p_{1,2} = \min\left\langle \Pr\left\{T \le t^{oss}\right\}, \Pr\left\{T \ge t^{oss}\right\}\right\rangle$$

mentre l'altro è calcolato esattamente come:

$$p_{2,1} = 1 - \min\left\langle \Pr\left\{T \le t^{oss}\right\}, \Pr\left\{T \ge t^{oss}\right\}\right\rangle.$$

Analogamente alla distinzione fatta per i due p-value, è stata fatta una necessaria distinzione tra la probabilità di accettazione di una e dell'altra ipotesi alternativa per i diversi valori della vera differenza tra i due trattamenti.

Quello che ne è risultato sono stati allora due vettori di valori descriventi le due diverse probabilità. I valori, calcolati anche qui come numero di volte sul totale delle replicazioni in cui il test presenta un p-value inferiore questa volta ad $\alpha/2$ (soglia di significatività per i test a due code), sono riportati nella tabella 8:

id	Shift	Probabilità H4	Probabilità H3
1	-1,00	0,0000	0,9979
2	-0,90	0,0000	0,9939
3	-0,80	0,0000	0,9775
4	-0,70	0,0000	0,9322
5	-0,60	0,0000	0,8439
6	-0,50	0,0000	0,7027
7	-0,45	0,0000	0,6112
8	-0,40	0,0000	0,5131
9	-0,35	0,0003	0,4120
10	-0,30	0,0009	0,3208
11	-0,25	0,0019	0,2347
12	-0,20	0,0043	0,1665
13	-0,15	0,0073	0,1067
14	-0,10	0,0170	0,0726
15	-0,05	0,0277	0,0439
16	0,00	0,0486	0,0254
17	0,05	0,0836	0,0149
18	0,10	0,1278	0,0057
19	0,15	0,1835	0,0035
20	0,20	0,2648	0,0019
21	0,25	0,3377	0,0008
22	0,30	0,4407	0,0000
23	0,35	0,5348	0,0001
24	0,40	0,6232	0,0000
25	0,45	0,7258	0,0000
26	0,50	0,7921	0,0000
27	0,60	0,9080	0,0000
28	0,70	0,9651	0,0000
29	0,80	0,9893	0,0000
30	0,90	0,9979	0,0000
31	1,00	0,9994	0,0000

Tabella 8: probabilità di accettazione rispettivamente delle ipotesi H_4 e H_3 per 31 diversi valori della vera differenza (shift) tra gli effetti dei due trattamenti, di controllo e sperimentale; le suddette probabilità sono state calcolate rispettivamente come numero totale dei p-value$_1$ (inteso come il p-value calcolato sulla coda destra, cioè quella relativa all'ipotesi alternativa H_4) inferiori a valore scelto per $\alpha/2=0.025$ sul totale dei 10000 p-value$_1$ calcolati per le altrettante replicazioni effettuate nello studio, e analogamente per il p-value$_2$.

Dai valori riportati in tabella possiamo vedere come per vere differenze tra gli effetti dei trattamenti minori di zero, cioè nel caso in cui è vera l'ipotesi alternativa H_3, la probabilità di accettazione di quest'ultima risulta essere molto alta, mentre alquanto ridotta è quella relativa all'accettazione dell'altra ipotesi alternativa; vale il viceversa per valori dello shift maggiori di zero, le cui probabilità associate sono riportate nella seconda parte della tabella.
Anche per questo studio riportiamo di seguito in figura 7 il grafico relativo ai valori descritti in tabella: in questo caso abbiamo voluto aggiungere la linea di colore rosso per indicare il livello $\alpha/2 = 0.025$ e la linea di colore blu per indicare il punto critico, lo zero appunto, a cavallo tra le due ipotesi alternative:

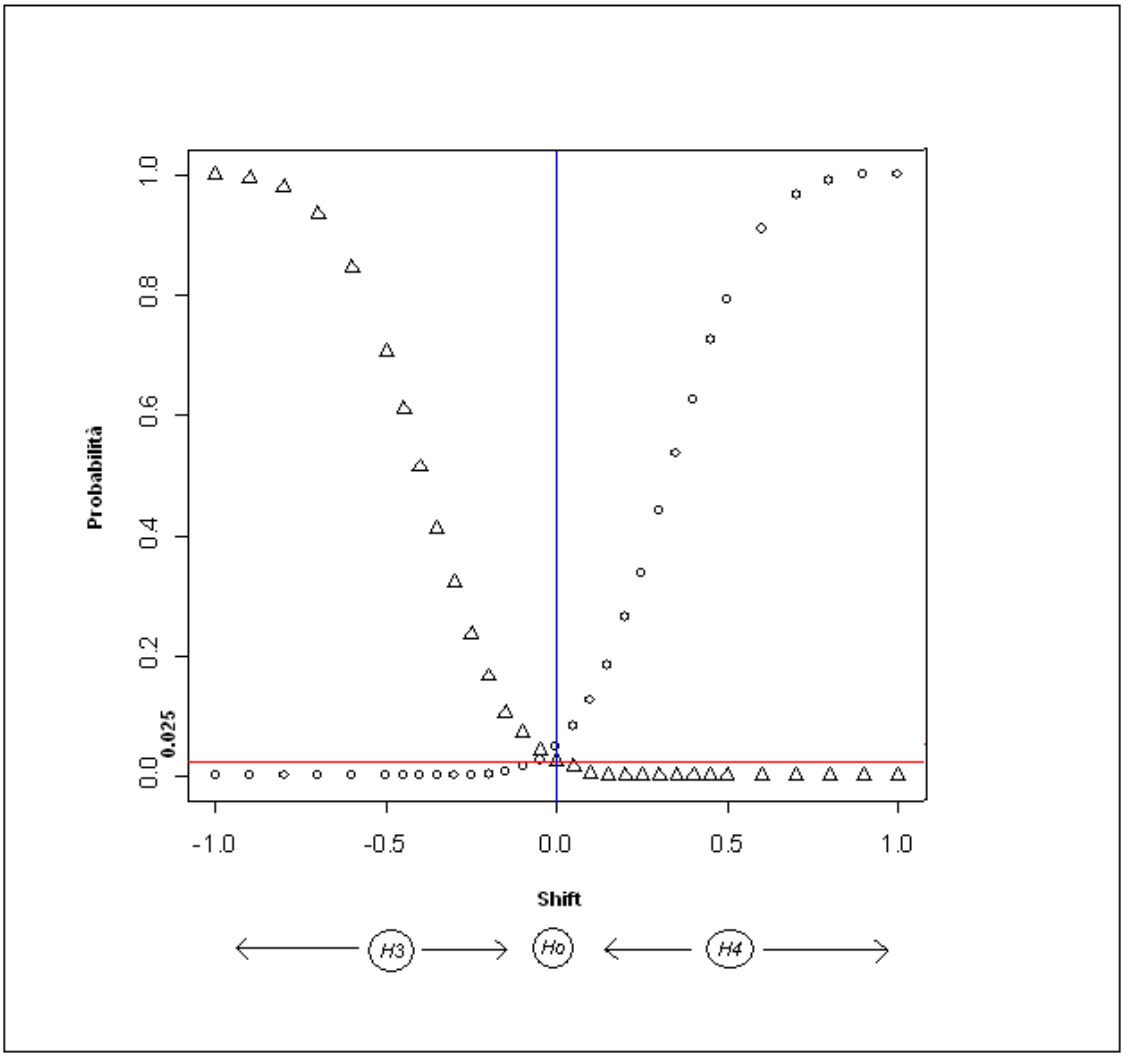


Figura 8: probabilità di accettazione rispettivamente delle ipotesi H_4 e H_3 per 31 diversi valori della vera differenza (shift) tra gli effetti dei due trattamenti, di controllo e sperimentale; le suddette probabilità sono state calcolate rispettivamente come numero totale dei p-value$_1$ (inteso come il p-value calcolato sulla coda destra, cioè quella relativa all'ipotesi alternativa H_4) inferiori a valore scelto per $\alpha/2=0.025$ sul totale dei 10000 p-value$_1$ calcolati per le altrettante replicazioni effettuate nello studio, e analogamente per il p-value$_2$.

2.3. GLI AMBITI DI UTILIZZO

2.3.1 Vantaggi e svantaggi del test direzionale a due code

Abbiamo fin qui analizzato il test direzionale a due code, formalizzando le sue tre ipotesi, definendo il concetto nuovo di errore del III tipo e confrontandolo inevitabilmente in termini di prestazioni (entità degli errori, potenza e numerosità campionarie richieste) con i test ad una o due code standard. Il quadro che ne è uscito è sicuramente un quadro che si presenta a favore del test qui analizzato, ricco di vantaggi rispetto agli altri test presi in considerazione.

Quest'ultimo paragrafo verrà allora dedicato esclusivamente al confronto tra i tre test, in modo tale da fornire un quadro chiaro delle differenze tra di essi e così da permettere di capire anche le situazioni pratiche in cui conviene utilizzare un test piuttosto che un altro. Si comincerà quest'analisi discutendo una critica che spesso viene fatta all'utilizzo dei test per prendere decisioni statistiche: quasi sempre accade che le ipotesi nulle dei test siano false, e questo comporta che l'applicazione dei test stessi divenga inutile, in quanto essa non ci porta a sapere nulla di nuovo; verrà mostrato come per i test direzionali ad una e due code questa critica riesca ad essere confutata.

In seguito saranno valutate le possibili tipologie di situazioni in cui conviene utilizzare il test direzionale a due code piuttosto che gli altri due test, o viceversa.

2.3.2 Una critica ai test

L'uso dei test per prendere decisioni statistiche è stato spesso anche in passato criticato, e come già accennato una critica in particolare è stata mossa[29]; in questo paragrafo dimostreremo come essa possa essere portata avanti solamente per il test non direzionale a due code, ma non per i test direzionali ad una e due code.

Due generazioni di ricercatori hanno affermato che quasi sempre le ipotesi nulle dei test statistici sono false; alcuni hanno argomentato la critica in modo più

[29] Leventhal L.; "Answering two criticisms of hypothesis testing", in *Psychological Reports*, (pp 3-6); 1999.

moderato, affermando che due popolazioni non sono mai esattamente uguali e che due popolazioni non possono mai avere correlazione uguale a zero. Altri più radicali affermano che specificare un valore per un certo parametro non sarà mai corretto perché il suddetto parametro non assumerà mai esattamente quel valore o una determinata curva di distribuzione teorica non sarà mai esattamente duplicabile da dati empirici.

Credo sia necessario esaminare la critica separatamente per i tre test presi in considerazione in questo capitolo per potersi rendere conto come essa possa essere portata avanti solamente per il primo test che verrà analizzato.

Per quanto riguarda il test non direzionale a due code effettivamente ci accorgiamo come sia inutile portare avanti l'analisi se già si conosce a priori che l'ipotesi nulla è falsa: se le due ipotesi (nulla ed alternativa) sono esaustive e mutuamente esclusive è chiaro che il fatto che l'ipotesi nulla sia certamente falsa fa si che sia automaticamente vera l'unica ipotesi alternativa.

Per quanto riguarda il test direzionale ad una coda, ricordiamo che esistono due diverse formalizzazioni dello stesso test: infatti l'ipotesi nulla del test di superiorità può venire a volte espressa in maniera imprecisa:

$$\begin{cases} H_2: & C \le T \\ H_3: & C > T \end{cases}$$

oppure in maniera precisa:

$$\begin{cases} H_2: & C = T \\ H_3: & C > T \end{cases}$$

Le regioni di rifiuto e di accettazione delle due formalizzazioni del test sono ovviamente le stesse, e l'ipotesi nulla $H_2: \quad C = T$ è semplicemente un caso particolare della prima versione $H_2: \quad C \le T$, che è evidentemente il modo più accurato e preciso di esprimere il test stesso. La presenza di questa prima versione imprecisa rende però inattaccabile il test dalla critica in oggetto in quanto l'ipotesi nulla in questo caso non descrive un singolo valore puntuale per il parametro oggetto di studio, ma un insieme di valori.

Anche il test direzionale a due code riesce a resistere alla critica in oggetto: esso infatti è costituito da tre ipotesi, una nulla e due alternative, tra loro come sappiamo esaustive e mutuamente esclusive, e questo comporta che anche se si sa già a priori che l'ipotesi nulla è falsa si dedurrà che è vera una delle due

ipotesi alternative, ma non si saprà quale delle due, e di conseguenza sarà necessaria l'applicazione del test per conoscerlo.
Ancora una volta quindi abbiamo conferma del passo in avanti fatto dal test direzionale a due code rispetto alla versione non direzionale, e ci rendiamo conto dell'efficienza del test di superiorità che può intercambiarsi con il test a tre ipotesi, qualora, come già detto, si abbia una certa sicurezza nell'ipotesi preliminare alla formulazione delle ipotesi del test stesso.

2.3.3 Una o due code?

Durante tutta la trattazione di questo capitolo sono stati ripetutamente messi a confronto tre test: il test non direzionale a due code, il test direzionale ad una coda e il test a tre ipotesi, come definiti nel primo paragrafo. Abbiamo confrontato le performance dei tre test, in termini di potenza, probabilità degli errori e ampiezze campionarie, e ci siamo fatti un'idea sul loro utilizzo ideale, e sulle scelte da fare da un punto di vista prettamente statistico. Credo che a questo punto sia necessario concludere con un confronto in termini di ambiti di utilizzo in un'ottica generale, in modo da chiarire pregi e difetti dei tre test nelle diverse situazioni di applicazione.
Partendo con il test non direzionale a due code, possiamo dire che il suo utilizzo reale si riduce a pochi casi in ambito di studi clinici per confrontare le prestazioni di due diversi trattamenti medici: quasi sempre infatti in questi campi risulta necessario indagare sulla direzione della differenza tra i due trattamenti analizzati, in quanto, come è facile intuire, di poca utilità è sapere che esiste una differenza di performance ma non sapere in che senso essa sia.
Il test direzionale ad una coda (o test di superiorità) ha molte più possibilità di applicazione, in quanto va ad indagare proprio la direzione delle differenza tra i due trattamenti; esso è più efficiente in termini di potenza, e quindi ha una probabilità di commettere un errore del II tipo minore rispetto ad un test a due code, però ha anche un grande svantaggio: esso necessita di ipotizzare una direzione della differenza oggetto di studio per formulare le ipotesi del test stesso, ma se viene ipotizzata la direzione sbagliata il test porterà a decisioni inaffidabili. Questo aspetto risulta alquanto delicato, data la delicatezza del

campo di applicazione del test stesso: possiamo affermare allora che il test direzionale ad una coda rimane applicabile mantenendo una certa sicurezza solamente nel caso in cui si abbia ragione di credere di conoscere con un margine di dubbio piuttosto basso la vera direzione della differenza di efficacia indagata.

Il test direzionale a due code rappresenta spesso la soluzione ideale per indagare sulla differenza di performance tra due trattamenti: esso è si meno potente del test di superiorità, ma in situazioni di incertezza è nettamente preferibile in quanto in grado di tenere in considerazione l'errore del III tipo definito come la probabilità di accettare una delle due ipotesi alternative quando invece è vera l'altra. Con questo test infatti si riesce ad indagare in entrambe le direzioni, cosa che spesso risulta desiderabile in campo medico.

Abbiamo visto nel corso del capitolo che il test direzionale a due code può essere visto come l'applicazione simultanea di due test direzionale ad una coda che indagano le due direzioni opposte; questo punto di vista è stato ripreso nello studio di simulazione, dove sono stati distinti due diversi p-value, per le due distinte regioni di accettazione delle due ipotesi alternative:

$$p_{1,2} = \min\left\langle \Pr\left\{T \le t^{oss}\right\}, \Pr\left\{T \ge t^{oss}\right\}\right\rangle$$

e

$$p_{2,1} = 1 - \min\left\langle \Pr\left\{T \le t^{oss}\right\}, \Pr\left\{T \ge t^{oss}\right\}\right\rangle.$$

che possono essere riscritti nel modo seguente, per cogliere meglio l'analogia dei singoli p-value con i p-value di due distinti test direzionale ad una coda[30]:

$$p^{+} = \Pr\left\{T \ge t^{oss}\right\}$$

e

$$p^{-} = \Pr\left\{T \le t^{oss}\right\}$$

La presenza di questa distinzione permette a mio avviso di cogliere ancora meglio la funzione del test direzionale a due code come duplice test ad una coda.

Abbiamo descritto come un risultato analogo possa essere raggiunto anche tramite l'applicazione di due test, prima un test non direzionale a due code per verificare l'eventuale significatività statistica della differenza indagata e poi, in

[30] Cox D. R., Hinkley D. V.; *Theoretical Statistics*, Chapman and Hall, London (p 79); 1974.

caso di accettazione dell'ipotesi alternativa, di un test direzionale ad una coda, impostando le ipotesi nel modo corretto a seconda del valore registrato per la statistica test della verifica di ipotesi precedentemente condotta.

In entrambi i modi alternativi di vedere il test direzionale a due code ci accorgiamo che esso svolge di fatto la funzione di due test distinti: è chiara allora a questo punto la convenienza di applicare un unico test, anche a discapito di una perdita di potenza, che sia in grado di calcolare la probabilità di tutti e tre i tipi di errori α, β e γ.

BIBLIOGRAFIA

1. Cox D. R., Hinkley D. V.; *Theoretical Statistics*, Chapman and Hall, London; 1974.

2. D'Agostino R. B.; "Non-inferiority trias: design concepts and issues – the encounters of academic consultants in statistics", in *Statistics in Medicine*, vol. 22, n. 2; 2003.

3. Dunnett C. W., Gent M.; "An alternative to use of two-sided tests in clinical trials", in *Statistics in Medicine*, vol. 15; 1996.

4. Hopkins B.; "Educational research and Type III errors", in *The Journal of Experimental Education*, vol. 41, n. 4; summer 1973.

5. Hung H. M. J., Wang S., Tsong Y., Lawrence J., O'Neil R. T.; "Some fundamental issues with non-inferiority testing in active controlled trials", in *Statistics in Medicine*, vol. 22, n. 2; 2003.

6. Kaiser H. F.; "Directional statistical decisions", in *Psychological Review*, vol. 67, n. 3; 1960.

7. Kimmel H. D.; "Tree criteria for the use of one-tailed tests", in *Psychological Bulletin*, vol. 54, n.4; 1957.

8. Laster L. L., Johnson M. F.; "Non-inferiority trials: the 'at least as good as' criterion"; in *Statistics in Medicine*, vol. 22, n. 2; 2003.

9. Leventhal L.; "Answering two criticisms of hypothesis testing", in *Psychological Reports*, vol. 85; 1999.

10. Leventhal L., Huynh C. L.; “Directional decisions for two-tailed tests: Power, Error Rates and Sample Size”, in *Psychological Methods*, vol. 1, n. 3; 1996.

11. Rashid M. M.; “Rank-based tests for non-inferiority and equivalence hypotesis in multi-centre clinical trials using mixed models”, in *Statistics in Medicine*, vol. 22, n. 2; 2003.

12. Shaffer J. P.; “Directional statistical hypothesis and comparisons among means”, in *Psychological Bulletin*, vol. 77, n. 3; 1972.

13. Wang S., Hung H. M. J.; “TACT method for non-inferiority testing in active controlled trials”, in *Statistics in Medicine*, vol. 22, n. 2; 2003.

14. EMEA – Committee for proprietary medicinal products; “Points to consider on the choice of non-inferiority margin”; London; 2004.