Università
Ca'Foscari
Venezia

Corso di Laurea Magistrale
in Management e Sostenibilità

Tesi di Laurea

Sulla potenza dei metodi di combinazione di p-value indipendenti

Relatrice

Ch.ma Prof.ssa Vera Djordjilovic

Correlatore

Ch. Prof. Aldo Solari

Laureando

Federico Toselli

Matricola 889508

Anno Accademico

2024/ 2025

"Comunque vada, sarà un successo."
– mio papà

Indice

Introduzione **1**

1 Test d'ipotesi e potenza: concetti essenziali **3**

1.1 Test d'ipotesi singoli . . . 3

1.1.1 La media di una normale . . . 4

1.1.2 Statistica test (σ^2 noto e non) . . . 6

1.1.3 Decisioni basate sulla regione di rifiuto e valore critico . . . 8

1.1.4 Decisioni basate sul *p-value* . . . 9

1.2 Errore di I° tipo . . . 11

1.3 Errore di II° tipo e potenza del test . . . 12

2 Metodi di aggregazione di p-value indipendenti **15**

2.1 Assunzioni alla base dei metodi di combinazione . . . 16

2.2 Metodo di Bonferroni . . . 18

2.3 Metodo di Šidák . . . 21

2.4 Metodo di Simes . . . 23

2.5 Metodo di Fisher . . . 24

2.6 Metodo di Stouffer . . . 26

3 Disegno dello studio di simulazione **29**

3.1 Domanda statistica operativa . . . 31

3.2 Ipotesi nulla ed alternativa . . . 31

3.3 Parametri della simulazione . . . 32

3.4 Modello di generazione dei dati e dei rispettivi *p-value* . . . 34

3.4.1 Modello di generazione dei dati . . . 34

3.4.2 Come vengono ottenuti i *p-value* . . . 35

3.5 Applicazione dei metodi combinatori . . . 36

4 Risultati dello studio di simulazione **39**

4.1 Scenario con m=2 . . . 39

4.2 Scenario con m>2 40
4.3 Risultato generale 44

5 Discussione, limiti e sviluppi futuri **45**
5.1 Discussione 45
5.2 Limiti e sviluppi futuri 47

Conclusioni **49**

A Comandi in R **51**
A.1 Disegno di simulazione con m=2 51
A.2 Disegno di simulazione con m>2 omogenea 53
A.3 Disegno di simulazione con m>2 lineare 55

B Risultati **59**
B.1 Tabella dei risultati con m=2 59
B.2 Tabelle dei risultati con m>2 omogenea 60
B.3 Tabelle dei risultati con m>2 lineare 62

Bibliografia **65**

Introduzione

Il test d'ipotesi rappresenta uno strumento fondamentale, alla base di gran parte delle scoperte statisticamente rilevanti. Proprio per questo è stato e continua ad essere oggetto di studio nei più diversi ambiti. Nello specifico, talvolta si è in presenza di test d'ipotesi multipli, ossia il caso in cui la ricerca si concentra su più ipotesi nulle da studiare. In questi casi, un modo di trattare la molteplicità dei test è quello di analizzare un'unica ipotesi nulla globale che consideri tutte le ipotesi nulle dei diversi test. Per mettere in atto tale approccio, sarà necessario combinare i diversi *p-value*, che si assumeranno essere indipendenti tra loro.

I metodi di combinazione di *p-value* indipendenti che si considereranno in questo studio sono quelli maggiormente conosciuti, ossia il metodo di Fisher, il metodo di Stouffer e il metodo di Bonferroni con i conseguenti metodi di Sidak e di Simes. Per cui, essendo diversi tra loro, si desidera comprendere quale sia il loro comportamento al variare di determinate caratteristiche del test d'ipotesi globale.

Nello specifico, l'obiettivo del presente elaborato è quello di analizzare come varia la potenza del test a seconda del metodo utilizzato per la combinazione dei *p-value* considerando la variazione della numerosità dei test, della dimensione dell'effetto e del livello di *sparsity*.

Quindi la tesi si propone di individuare quale (o quali) metodo garantisca la potenza maggiore nelle specifiche situazioni, contribuendo quindi a quel filone della letteratura che studia (anche più approfonditamente) i diversi metodi di combinazione dei *p-value* e le loro applicazioni nel contesto dei test d'ipotesi multipli.

Nella parte iniziale dell'elaborato si presenteranno i fondamenti della statistica di base che verranno utilizzati nel corso dello studio. Successivamente verranno illustrati i metodi di combinazione più utilizzati in ambito di test d'ipotesi multipli, focalizzandosi anche sulla teoria sottostante ai medesimi. Dopodiché, si descriverà il disegno dello studio di simulazione, con particolare attenzione al modello scelto ed alle assunzioni formulate. Una volta effettuata la simulazione attraverso il codice R (Appendice A), si riporteranno i risultati più rilevanti con delle osservazioni utili al

perseguimento dell'obiettivo. Infine, nell'ultima parte verrà discusso quanto ottenuto e verranno evidenziati anche i limiti dello studio di simulazione svolto, oltre a possibili sviluppi futuri.

Capitolo 1

Test d'ipotesi e potenza: concetti essenziali

Per perseguire l'obiettivo del presente elaborato, ossia lo studio della potenza statistica nei metodi di combinazione di *p-value* indipendenti, che sono anche alla base dei metodi di verifica della replicabilità di un risultato scientifico, sarà primariamente necessario introdurre alcuni concetti base che verranno usati successivamente.

Partendo dal presupposto che si abbia un minimo di conoscenza e di dimestichezza con gli elementi della statistica elementare [vedi ad es. Ross (2010), in particolare fino pag. 330], si presenterà la verifica d'ipotesi focalizzandosi sugli strumenti alla base dell'inferenza come la statistica test, il concetto di regione di rifiuto e di *p-value*, analizzando anche i due tipi di errori che possono essere commessi qualora la conclusione dello studio fosse sbagliata.

1.1 Test d'ipotesi singoli

La verifica d'ipotesi è il metodo statistico che permette di decidere sulla plausibilità (o meno) di un'ipotesi riguardante uno specifico parametro θ, utilizzando dei dati campionari che permettano di prendere una decisione [Agresti et al. (2017), pp.386 ss.]. All'interno di tale definizione è possibile evidenziare tutti gli elementi essenziali per un test d'ipotesi, ovvero il concetto di parametro (θ), ossia la quantità incognita della popolazione oggetto di analisi, sulla quale viene formulata un'ipotesi detta nulla (H_0), che generalmente esprime assenza di effetto o di differenze. A questa si contrappone l'ipotesi alternativa (H_1), che, qualora l'evidenza empirica porti al rifiuto di H_0, consente di concludere per la presenza di un effetto o di un risultato statisticamente significativo.

Per semplificare la spiegazione si presenta di seguito un esempio pratico (e quanto più semplice possibile).

Ipotizzando di voler sapere se il tempo medio di ricarica (parametro oggetto dell'ipotesi) per un determinato modello di telefono è sempre inferiore alle 8 ore, l'ipotesi nulla dovrà cercare di dimostrare l'opposto rispetto a quanto ipotizzato (caso che non porterebbe ad una conclusione). Per cui, dato θ uguale al tempo medio di ricarica totale del telefono modello xyz e θ_0 uguale al valore di riferimento (8 ore), il test d'ipotesi si presenterebbe come nella Formula 1.1:

$$\begin{cases} H_0 : \theta \geq \theta_0 \\ H_1 : \theta < \theta_0 \end{cases} \tag{1.1}$$

Tale illustrazione può essere utilizzata in qualunque test d'ipotesi purché venga cambiato il significato di θ ed il corrispondente valore di riferimento, oltreché cambiare il segno qualora necessario (di solito per $H_0 \leq, \geq, =$).

Inoltre, è d'obbligo introdurre la distinzione tra ipotesi nulla semplice e composta [Lehmann & Romano (2005)]. Un'ipotesi nulla viene definita semplice se essa specifica completamente la distribuzione della popolazione. Mentre, altrimenti, l'ipotesi nulla si dice composta [Rossi (2018), p. 281].

In questo paragrafo si cercherà di illustrare i metodi e gli strumenti che è possibile utilizzare per compiere effettivamente il test delle ipotesi. Inizialmente saranno presentate le diverse casistiche di ipotesi, per passare successivamente al calcolo della statistica test e le conseguenti decisioni che possono essere prese. Infine, si analizzeranno i tipi di errori che possono essere commessi quando viene tratta una conclusione sbagliata.

1.1.1 La media di una normale

Spesso il parametro che viene preso in considerazione per il test d'ipotesi è la media (μ) della popolazione da cui è stato estratto il campione di riferimento per svolgere il test. In questi casi, la varianza (σ^2) della popolazione può essere nota o meno. Come è intuibile, si illustreranno inizialmente i casi di test su medie con varianza nota e successivamente si procederà con una breve presentazione dell'approccio da utilizzare quando la varianza non è nota, il quale comporta principalmente una modifica della statistica test, per cui verrà trattato nel sottoparagrafo relativo a quest'ultima.

Se si volessero collegare due strumenti della statistica inferenziale, si potrebbe affermare che tali tipologie di test avvengono attraverso una stima intervallare che permette di affermare con un certo grado di confidenza $(1 - \alpha)$ che la media si trova all'interno di uno specifico intervallo [per approfondire vedi Cicchitelli et al. (2021),

pp. 383 ss.;Anderson et al. (2019), pp. 434 ss.; però cfr. Casella & Berger (2002), p. 373].

A seconda di ciò che si vuole verificare e quindi che termine di confronto avrà l'ipotesi nulla, si possono ottenere tre tipi di test. Ovvero test d'ipotesi ad alternativa destra (che individua la regione critica nella coda destra, dove si collocano i valori a favore dell'ipotesi alternativa), sinistra o bilaterale.

Alternativa unilaterale

Qualora il test d'ipotesi sulla media della popolazione fosse come nella formula riportata di seguito:

$$\begin{cases} H_0 : \mu \leq \mu_0 \\ H_1 : \mu > \mu_0 \end{cases}$$

Allora si avrebbe un test d'ipotesi ad alternativa (o a coda) destra, in cui, per l'appunto, la regione di rifiuto del test si trova a destra. Ed ipotizzando una statistica test Z con distribuzione normale standard sotto l'ipotesi nulla ($Z \sim N(0,1)$), nel quale z=$\frac{\overline{x}-\mu_0}{\sigma/\sqrt{n}}$ rappresenta la statistica test osservata e, di conseguenza, l'area colorata è il *p-value* sulla base dei quali andranno prese le decisioni; si otterrà la Figura 1.1:

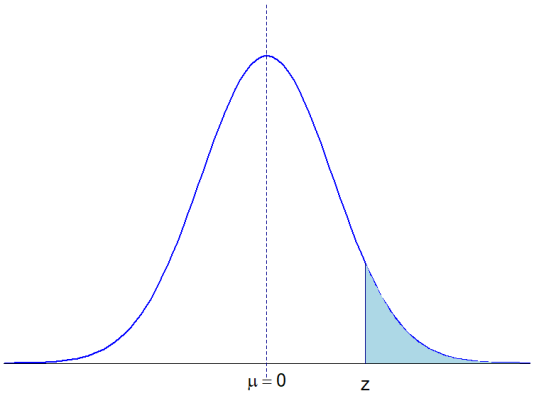

Figura 1.1: Alternativa Destra in una Normale Standard

Per simmetria, quindi se il test d'ipotesi fosse ad alternativa (o a coda) sinistra, ovvero:

$$\begin{cases} H_0 : \mu \geq \mu_0 \\ H_1 : \mu < \mu_0 \end{cases}$$

Allora il *p-value* sarà rappresentato dall'area sotto la curva alla sinistra del valore osservato di z nella Figura 1.1

Alternativa bilaterale

Nel caso in cui il test d'ipotesi sulla media della popolazione fosse del tipo:

$$\begin{cases} H_0 : \mu = \mu_0 \\ H_1 : \mu \neq \mu_0 \end{cases}$$

Il test bilaterale può, quindi, essere interpretato come l'unione dei due test unilaterali, uno a destra e uno a sinistra, e il *p-value* si calcola come il doppio del minore tra i due *p-value* unilaterali. Da cui ne conseguirà la Figura 1.2:

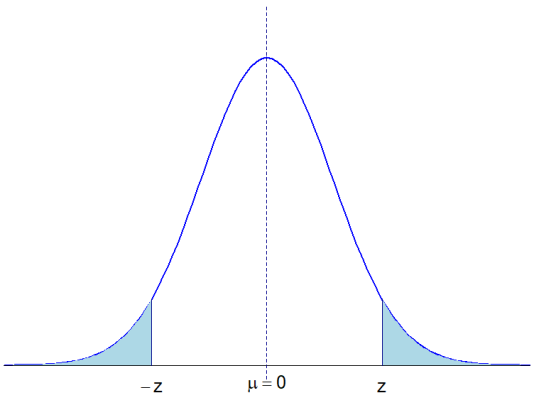

Figura 1.2: Alternativa bilaterale in una Normale Standard

1.1.2 Statistica test (σ^2 noto e non)

Dopo aver analizzato le tre possibili casistiche in tema di ipotesi su μ, è necessario esplicitare il calcolo della statistica test che potrebbe essere visto come quel valore che, come si è già accennato, permette di comprendere quando H_0 può essere rifiutata e quando non è possibile[1]. La formula generale della statistica test nel caso σ^2 sia noto è [Hogg et al. (2005), p. 268]:

$$Z = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} .$$

Dove $\overline{X}$ è la variabile casuale rappresentante la media campionaria di riferimento, μ_0 è il valore ipotizzato della media della popolazione, σ è la deviazione standard della popolazione (in questo caso nota) e n è il numero di osservazioni raccolte nel campione. Per cui, riprendendo il Teorema del limite centrale (TLC) il quale afferma che, per un numero elevato di osservazioni (nella pratica per $n > 30$), la distribuzione

[1] vedi par. 1.1.3

è approssimabile ad una normale, la formula della statistica test appare essere sostanzialmente una standardizzazione della media campionaria rispetto a quella di riferimento (μ_0) in base alla deviazione standard della popolazione che, essendo divisa per $\sqrt{n}$, rappresenta la deviazione standard della media campionaria[Clarke & Cooke (1992), pp. 233-235].

In particolare, la formula è (ovviamente) in linea con le regole logiche più elementari. Infatti, se aumentano le osservazioni, diminuirà l'impatto della deviazione standard della popolazione sulla differenza tra media campionaria ed ipotizzata, per cui tale gap verrà talvolta (se $0 < \sigma/\sqrt{n} < 1$) amplificato dalla presenza di molte osservazioni. Ciò è dovuto dal fatto che più osservazioni si svolgono più il campione tenderà ad eguagliare la popolazione, ragionamento che sottende al TLC, per cui una differenza tra ipotesi e campione segnalerà una problematica nell'ipotesi stessa.

Mentre, a parità di differenza tra medie e di numero di osservazioni, tanto maggiore è la deviazione standard della popolazione tanto minore sarà la statistica test e quindi il grado di rifiutabilità di H_0 sarà maggiore, come si vedrà nel prossimo sottoparagrafo.

È ora necessario analizzare il caso in cui la varianza (e di conseguenza σ) non fosse nota. Dato che gli unici due fattori che cambiano sono la statistica test e la distribuzione di riferimento, si tratterà di questa casistica solo nel presente sottoparagrafo, sapendo però che i metodi e le riflessioni che si faranno di seguito sono parimenti valide purché vengano cambiati i due elementi distintivi di cui si è appena detto.

La statistica test, nel caso di σ^2 non noto, prende come riferimento s (deviazione standard campionaria) al posto di quella della popolazione. Facendo tale modifica, è necessario modificare anche la distribuzione di riferimento che passa ad una t di Student con $n-1$ gradi di libertà, $T \sim t_{n-1}$ [Newbold et al. (2013) , pp. 362 ss.]. Per cui la formula per la statistica test osservata che ne consegue è la seguente:

$$t_{obs} = \frac{\overline{x} - \mu_0}{s/\sqrt{n}} \, .$$

Nel prosieguo dell'elaborato si considererà il caso di varianza nota e quindi come statistica test di riferimento: z con distribuzione $Z \sim N(0,1)$. Tenendo però a mente che tali strumenti possono essere usati anche nel caso si utilizzino altre statistiche test, prendendo atto delle dovute accortezze.

1.1.3 Decisioni basate sulla regione di rifiuto e valore critico

Nel sottoparagrafo precedente si è detto come la statistica test possa essere intesa come quel punto che permette il rifiuto o meno dell'ipotesi nulla (H_0) ed il primo metodo che si analizzerà sfrutta proprio questo suo ruolo per arrivare ad una decisione.

Occorre però primariamente definire esattamente cosa significa "Rifiutare H_0". Infatti, si è visto sopra come l'ipotesi nulla sia quella che non produce effetti/conclusioni, per cui per giungere ad un risultato appare logico che tale ipotesi debba essere rifiutata. Poiché, qualora questo fosse possibile, si potrà ripiegare verso l'ipotesi alternativa, la quale invece produce l'effetto desiderato, ovvero non rifiutare l'ipotesi alla base del test con un certo grado di significatività che viene definito inizialmente.

Le decisioni basate sulla regione di rifiuto richiedono preliminarmente l'individuazione del punto (o valore) critico, ovvero di quel punto che delimita la zona di rifiuto di H_0 rispetto a quella in cui non è possibile rifiutarla per non incorrere nell'errore di I° tipo[2]. Il punto critico dipende dal grado di significatività che si desidera avere nel test di ipotesi, questo livello rappresenta la probabilità massima tollerata dal test di commettere l'errore di I° tipo. Solitamente il grado di riferimento maggiormente utilizzato negli studi statistici è 0.05, anche se alcuni studiosi [Benjamin et al. (2018)] nel recente passato hanno proposto di abbassare tale livello a 0.005 in modo da dare più credibilità alle nuove scoperte.

Assumendo un grado di significatività $\alpha = 0.05$ è necessario individuare il valore critico che nella distribuzione normale standard corrisponde ad una regione di rifiuto pari a 0.05 e, quindi, dove la probabilità cumulata è pari a 0.95. Nel caso specifico il valore critico è pari a 1.65, in caso di un test con coda destra.

Ora è necessario confrontare tale valore con la statistica test osservata, per comprendere se quest'ultima rientri o meno nella regione di rifiuto dell'ipotesi nulla. Mantenendo un grado di significatività di $\alpha = 0.05$ e quindi un punto critico pari a 1.65 (o -1.65 in caso di alternativa sinistra) si potrebbe semplificare il ragionamento da fare con la Figura 1.3.

Per cui qualora la statistica test osservata rientrasse nella regione di rifiuto dell'ipotesi nulla (z_1) si potrà giungere alla conferma dell'ipotesi alla base del test mediante il rifiuto di H_0; altrimenti (z_2) col campione analizzato non si sarà in grado di affermare ciò. Quindi, si dovrà rivedere l'ipotesi iniziale, il campione e/o il grado di significatività.

[2] vedi par. 1.2

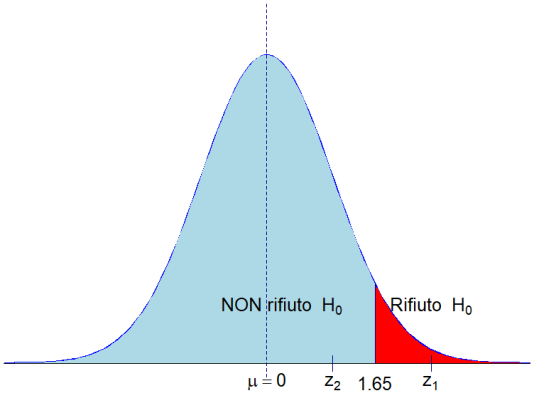

Figura 1.3: Decisioni basate sulla regione di rifiuto

Nel caso in cui il test d'ipotesi avesse una coda a sinistra allora sarebbe necessario considerare l'opposto del punto critico di riferimento, quindi o selezionare il punto come si è appena illustrato e cambiare il segno oppure selezionare direttamente il punto in cui la probabilità cumulata è uguale ad α.

Qualora, invece, il test d'ipotesi fosse a due code (alternativa bilaterale) bisognerebbe considerare come livello di significatività, da utilizzare per trovare il punto critico, la metà di α. Quindi riprendendo il caso con $\alpha = 0.05$ sarà necessario trovare il punto critico che abbia un livello di significatività di $\alpha/2 = 0.025$, ovvero che abbia una probabilità cumulata di 0.975. Il valore critico in tal caso sarebbe 1.96, che però va considerato assieme al suo opposto per individuare le due aree (code) di rifiuto. Si potrebbe perciò riassumere nel dire che con un test a due code e un $\alpha = 0.05$, il test da svolgere per rifiutare l'ipotesi nulla è che z_{obs} sia o maggiore di 1.96 o minore di -1.96.

1.1.4 Decisioni basate sul *p-value*

Un secondo approccio che è possibile utilizzare per svolgere il test d'ipotesi è quello che si basa sul *p-value*, il quale è la probabilità, quando l'ipotesi nulla è vera, di ottenere un valore della statistica test almeno tanto estremo quanto quello che si è ottenuto dal campione analizzato [Lee et al. (2013), pp. 495-496].

Si potrebbe anche intendere tale valore come il livello di significatività minimo in cui H_0 può essere rifiutata [Noether (1991), p.97].

Per prendere delle decisioni basandosi sul *p-value* ovviamente andrà primariamente calcolato tale valore e successivamente andrà confrontato col livello di significatività (α) assunto nel test d'ipotesi. Per individuare il valore del *p-value* bisognerà calcolare l'area della probabilità cumulata nella normale standard che si trova dopo (o prima)

della statistica test osservata. Ovvero, in un test d'ipotesi ad alternativa destra, bisognerà calcolare la probabilità cumulata a destra della statistica test, mentre, per le alternative sinistre sarà necessario considerare l'area a sinistra di z e per quelle bilaterali bisognerà moltiplicare per due l'area più piccola tra quella a sinistra e quella a destra del valore osservato di z. Se si volesse dare una formula esplicita al *p-value* in caso di ipotesi alternativa unilaterale destra, esplicitando gli elementi essenziali visti nella definizione, questa sarebbe [Newbold et al. (2013)]:

$$p - value = P\left(\frac{\overline{X}-\mu_0}{\sigma/\sqrt{n}} \geq z_{obs} | H_0 : \mu = \mu_0\right) = P(N(0,1) \geq z_{obs})$$

$$dove\ z_{obs} = \frac{\overline{x}-\mu_0}{\sigma/\sqrt{n}}\ e$$

$$Z = \frac{\overline{X}-\mu_0}{\sigma/\sqrt{n}}\ \textit{è la statistica test, che sotto } \theta \in \Theta_0 \textit{ si distribuisce come N(0,1)}.$$

Una volta calcolato il *p-value*, lo si confronta col valore di α del test d'ipotesi. Tale confronto, a seguito di quanto è stato detto finora, non può che portare alla regola riassunta nella presente formula.

$$\text{Rifiuto } H_0 \text{ se: } p\text{-value} < \alpha \quad .$$

Alcuni hanno criticato l'uso del test di significatività, utilizzando anche il *p-value*. Infatti secondo questi studiosi potrebbe essere rischioso basare le decisioni riguardo ad uno (o più) studio solamente su un limite arbitrario come α e confrontarlo col *p-value* che può assumere valori anche molto prossimi al valore limite α, sia da indurre a rifiutare H_0 che a non rifiutarla. In particolare, accusano proprio questo approccio della presenza di errori di I° o II° tipo nelle ricerche [per approfondire tale critica, vedi Rothman (2010)].

Nel criticare il fatto di basare le decisioni su α e *p-value* non sono mancati alcuni spunti per possibili soluzioni. Per l'appunto, si è ipotizzato di poter usare una stima intervallare con un certo grado di confidenza, così da dare maggiore rilevanza ai dati raccolti piuttosto che ad una decisione dicotomica (Rifiuto o meno H_0) [Stang et al. (2010)]. Posto che l'utilizzo di uno metodo non preclude l'utilizzo dell'altro, a parere di chi scrive, si potrebbe affermare che le conclusioni basate sul test di significatività sono valide fin tanto che non si rientri in quelle situazioni in cui il valore ottenuto (*p-value* o statistica test) è prossimo al valore limite di riferimento (α o valore critico).

Allora, in tal caso, potrebbe essere utile stimare un intervallo di confidenza, in modo da dare maggiore evidenza alla decisione presa.

In conclusione al presente sottoparagrafo dedicato all'uso del *p-value* nei test d'ipotesi, è d'obbligo puntualizzare che: *"the p-value is not the probability that null hypothesis is true"*[Yoccoz (1991)].

1.2 Errore di I° tipo

Nel corso di questa prima parte dell'elaborato, si è più volte fatto riferimento al concetto di errore di I° tipo senza però darne una definizione precisa.

La definizione di questo tipo di errore è relativamente banale, ovvero si ha l'errore di I° tipo se si rifiuta H_0 quando questa è vera[Mukhopadhyay (2000), p. 396]. La probabilità di commettere tale tipo di errore viene indicata con α che, non a caso, corrisponde al grado di significatività dei test d'ipotesi. Infatti, come è già stato accennato, il livello di significatività altro non è che il grado massimo tollerabile di incorrere nell'errore di I° tipo.

Si ponga $H_0 : \theta \in \Theta_0$, con Θ_0 uguale ad una serie di valori che il parametro deve assumere perché l'ipotesi nulla sia vera [Olive (2014), p.183]. Allora la probabilità di commettere l'errore di I° tipo sarebbe calcolabile come segue:

$$\alpha = P\left(\text{Rifiutare } H_0 \mid \theta \in \Theta_0\right) .$$

L'importanza dell'errore di I° tipo appare lampante. Infatti, come si è più volte detto, solo con il rifiuto dell'ipotesi nulla è possibile giungere ad una conclusione, altrimenti se non la si rifiutasse non ne scaturirebbe alcun effetto. Nello specifico, ipotizzando di voler testare l'efficacia o meno di un determinato farmaco (esempio classico in tema di test d'ipotesi), si potrebbe svolgere un test d'ipotesi che indichi come a seguito dell'assunzione del farmaco il parametro θ (ad es. il livello di colesterolo) si abbassi rispetto a prima dell'assunzione (θ_0). Quindi il test assumerebbe la forma della Formula 1.1 che di seguito si riprende:

$$\begin{cases} H_0 : \theta \geq \theta_0 \\ H_1 : \theta < \theta_0 \end{cases}$$

Per cui, qualora si rifiutasse H_0 si arriverebbe alla conclusione che il farmaco sia efficace e quindi sarebbe commercializzato come tale. Però qualora si scoprisse che tale farmaco nel concreto non espleta la funzione che si propone di svolgere,

allora l'ipotesi nulla (inizialmente rifiutata) risulterebbe vera e quindi non si sarebbe dovuto rifiutarla, evitando così la commercializzazione del farmaco.

L'esempio appena svolto, serve a comprendere il motivo per cui viene sempre prioritariamente considerato l'errore di I° tipo nel prendere decisioni riguardo ad un test d'ipotesi, in cui viene preso appositamente a riferimento il livello di significatività α.

1.3 Errore di II° tipo e potenza del test

Oltre a rifiutare un'ipotesi nulla vera, potrebbe anche accadere di non rifiutarne una falsa, in tal caso si verificherebbe l'errore di II° tipo [Banerjee et al. (2009)]. Per cui, se si volesse riprendere la notazione di prima e ponendo che Θ_1 è l'insieme di valori di θ in cui l'ipotesi nulla è falsa (per cui $\Theta_1 \cup \Theta_0 = \Theta_i \subset \Theta$ dove Θ è l'insieme di tutti i possibili valori assumibili da θ [Olive (2014)]), si potrebbe affermare che la probabilità di commettere un errore di II° tipo è rappresentabile con:

$$\beta = P\left(\text{Non rifiutare } H_0 \mid \theta \in \Theta_1\right) .$$

Tale probabilità viene indicata con β e risulta essere particolarmente utile per comprendere la "bontà" del test d'ipotesi. Infatti, il complementare di β (che non è α) rappresenta la potenza del test, ovvero $1 - \beta$. Tale formula rappresenta la probabilità di rifiutare l'ipotesi nulla quando questa è falsa [Clarke & Cooke (1992), pp. 184 ss.], per cui potrebbe essere illustrata come:

$$1 - \beta = P\left(\text{Rifiutare } H_0 \mid \theta \in \Theta_1\right) .$$

Dato che il concetto di α è già stato ampiamente definito, è necessario ora esplicitare anche il calcolo di β [Newbold et al. (2013), p.369].

Sapendo come trovare il valore critico per uno specifico livello di significatività ($z_{1-\alpha}$), è possibile utilizzare tale valore per estrapolare $\overline{x}$ all'interno della regola alla base del rifiuto di H_0[3] ottenendo quindi, per il caso di mancato rifiuto (ossia con segno inverso):

$$\overline{x} < \mu_0 + z_{1-\alpha}(\sigma/\sqrt{n}) .$$

[3] Rifiuto H_0 se $\dfrac{\overline{x} - \mu_0}{\sigma/\sqrt{n}} > z_{1-\alpha}$.

Tale esplicitazione permette di individuare i valori della media campionaria nei quali H_0 viene rifiutata, limite[4] che verrà indicato con $\overline{x}_c$. Tuttavia è necessario comprendere quando l'ipotesi nulla non viene rifiutata anche se falsa. Quindi, affinché H_0 sia falsa, inevitabilmente H_1 deve essere vera, per cui va selezionato un valore del parametro di riferimento (in questo caso μ) che rispecchi tale ragionamento.

Quindi, prendendo come base un test d'ipotesi a coda destra e considerando μ_1 pari ad un qualsiasi valore di μ che renda falsa H_0, si assuma $\mu = \mu_1$ all'interno del calcolo di β come segue:

$$\beta = P(\overline{x} < \overline{x}_c \mid \mu = \mu_1) = P\left(Z < \frac{\overline{x}_c - \mu_1}{\sigma/\sqrt{n}} \mid \mu = \mu_1\right) \; dove \; Z = \frac{\overline{X} - \mu_1}{\sigma/\sqrt{n}} \sim N(0,1)\,.$$

Col fine di generalizzare la definizione fin qui data di Errore di II° tipo, si considera ora un qualunque parametro θ come oggetto d'ipotesi.

$$\beta = P\left(Z < \frac{\overline{x}_c - \theta}{\sigma/\sqrt{n}} \mid \theta = \theta_1\right) = P_{\theta_1}\left(Z < \frac{\overline{x}_c - \theta_1}{\sigma/\sqrt{n}}\right)\,.$$

Esplicitando la v.c. della statistica $Z = \frac{\overline{X} - \theta_1}{\sigma/\sqrt{n}}$ e sfruttando la sua distribuzione sotto H_1 ($Z \sim N(0,1)$) si giunge a

$$\beta = P_{\theta_1}\left(\frac{\overline{X} - \theta_1}{\sigma/\sqrt{n}} < \frac{\overline{x}_c - \theta_1}{\sigma/\sqrt{n}}\right) = \Phi\left(\frac{\overline{x}_c - \theta_1}{\sigma/\sqrt{n}}\right)\,.$$

Così come è possibile il calcolo nel caso di ipotesi alternativa destra, è possibile calcolare β anche in caso di ipotesi alternativa sinistra e bilaterale, purché si modifichi il termine di confronto (considerando rispettivamente la coda sinistra o entrambe le code).

A numerosità campionaria fissata, tra le due tipologie di errori c'è una relazione inversa per cui all'aumentare di α , β si ridurrà e viceversa [Kaur & Stoltzfus (2017)]. Questo è facilmente intuibile sia a livello logico che a livello di formule di calcolo, anche se a livello logico potrebbe essere non così immediato il ragionamento per cui riducendo la probabilità di commettere un errore di I° tipo aumenti quella del II° tipo, essendo due concetti diversi. Invece, a livello matematico (e di formule), è abbastanza semplice comprendere come se si decidesse di abbassare il livello di significatività sarebbe meno probabile rifiutare H_0 ed inoltre si alzerebbe il valore di $z_{1-\alpha}$ che a sua volta alzerebbe il valore di $\overline{x}_c$, incidendo sul termine di confronto con la statistica test osservata[5] e quindi aumentando la probabilità dell'errore di II° tipo.

[4] Generalizzando a θ, è pari a $\theta_0 + z_{1-\alpha}(\sigma/\sqrt{n})$

[5] $\left(z_{obs} < \frac{\overline{x}_c - \mu_1}{\sigma/\sqrt{n}}\right)$

Il che va a condizionare anche la potenza del test che appare così svilita. Per cui si potrebbe affermare che tra α e la potenza del test c'è una relazione diretta, mentre con β vi è ovviamente una relazione inversa.

Se si volesse definire matematicamente il concetto di potenza si potrebbe utilizzare quanto fatto finora per il calcolo di β e giungere ad una definizione del suo complementare, ossia $1-\beta$. Anche in questa circostanza, il focus sarà sul caso di ipotesi alternativa destra.

$$Potenza = 1 - P\left(Z < \frac{\overline{x}_c - \theta}{\sigma/\sqrt{n}} \mid \theta = \theta_1\right) = 1 - \Phi\left(\frac{\overline{x}_c - \theta_1}{\sigma/\sqrt{n}}\right) .$$

Un ulteriore passaggio che si potrebbe svolgere sia per la potenza che per β è quello di sostituire $\overline{x}_c = \theta_0 + z_{1-\alpha}(\sigma/\sqrt{n})$. Di seguito si riporta solamente il calcolo della potenza, la quale sarà l'argomento principale del presente elaborato.

$$Potenza = 1 - \Phi\left(\frac{\theta_0 + z_{1-\alpha}(\sigma/\sqrt{n}) - \theta_1}{\sigma/\sqrt{n}}\right) = 1 - \Phi\left(z_{1-\alpha} + \frac{\theta_0 - \theta_1}{\sigma/\sqrt{n}}\right)$$

In conclusione di questo capitolo introduttivo si desidera presentare, nella Figura 1.4, una rappresentazione dei concetti di errore di I° e II° tipo e di potenza, applicati al caso di un test a coda destra.

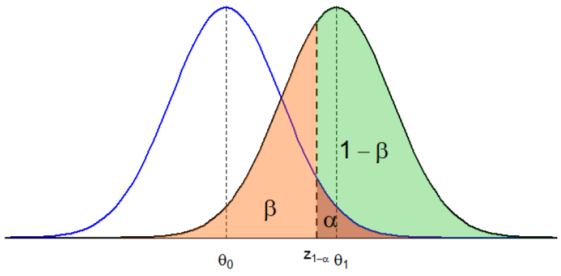

Figura 1.4: Rappresentazione grafica di potenza ed errore di I° e II° tipo

Capitolo 2

Metodi di aggregazione di p-value indipendenti

Nel capitolo precedente è stato presentato il test d'ipotesi nel caso di una singola ipotesi nulla; tuttavia, nelle applicazioni reali è frequente dover affrontare simultaneamente più ipotesi. A ciascuna di esse si può associare un *p-value*; l'obiettivo di questa tesi è quindi combinare tali *p-value* per verificare l'ipotesi nulla globale, ossia l'ipotesi secondo cui tutte le ipotesi nulle individuali sono vere. A tal fine verranno esaminati metodi di combinazione dei *p-value*, assumendo che essi siano indipendenti e uniformemente distribuiti sotto l'ipotesi nulla globale.

Anche in questo caso, l'ambito in cui tali procedure risultano particolarmente rilevanti è quello medico e biomedico. In particolare, vengono formulate simultaneamente molteplici ipotesi nulle con l'obiettivo di individuare quali variabili presentino variazioni significative a seguito di un esperimento di tipo *microarray*.

In tali esperimenti, infatti, si ottiene un numero molto elevato di misurazioni, spesso corrispondenti ai livelli di espressione di geni, che devono essere analizzate congiuntamente. Qualora si voglia effettuare un'inferenza statistica sulle medie delle singole variabili, ad esempio per valutare l'effetto di un trattamento sanitario rispetto a una condizione di controllo, è necessario formulare e testare simultaneamente un gran numero di ipotesi nulle, una per ciascuna variabile considerata.

Riprendendo la formulazione vista in precedenza, in cui si considerava l'ipotesi di uguaglianza tra la media μ e un valore prefissato μ_0, nel caso di ipotesi nulle multiple è sufficiente indicare esplicitamente la variabile i a cui si riferisce ciascuna ipotesi nulla. Più in generale, considerando m variabili, ciascuna ipotesi nulla può essere formulata come

$$H_{0,i} : \mu_i = \mu_{0,i} \, ,$$

[Bannerman-Thompson et al. (2012), pp. 201-238], dove $\mu_{0,i}$ rappresenta il valore di riferimento che si intende testare per la i-esima variabile (che, senza perdita di generalità, può essere posto pari a 0).

In questo contesto, testare preliminarmente l'ipotesi nulla globale consente di verificare se esista un'evidenza complessiva contro l'ipotesi che tutte le ipotesi individuali siano vere, evitando di procedere immediatamente con una molteplicità di test separati. Tale strategia permette di contenere il rischio di errori di I tipo dovuti alla molteplicità delle ipotesi e giustifica, solo in presenza di un risultato significativo a livello globale, un'analisi più approfondita delle singole ipotesi.

L'ipotesi nulla globale stabilisce che nessuna delle ipotesi nulle individuali sia falsa [Futschik et al. (2019); Lehmann & Romano (2005)], ossia che tutte le $H_{0,i}$ siano vere. Il rifiuto dell'ipotesi globale implica quindi che almeno una delle ipotesi nulle individuali sia falsa. Pertanto, per rigettare l'ipotesi nulla globale non è necessario che tutte le ipotesi individuali siano false, ma è sufficiente che ne esista almeno una falsa.

L'ipotesi nulla globale è definita come l'intersezione di tutte le ipotesi nulle individuali [Futschik et al. (2019)], ossia

$$H_0 = \bigcap_{i=1}^{m} H_{0,i} = \bigcap_{i=1}^{m} \{\mu_i = 0\} \,.$$

Essa può essere espressa in modo equivalente come

$$H_0 : \mu = 0 \,,$$

dove $\mu = (\mu_1, \mu_2, \ldots, \mu_m)$ rappresenta il vettore delle medie, cioè l'ipotesi secondo cui tutti i parametri μ_i sono simultaneamente nulli [Liu et al. (2024)].

2.1 Assunzioni alla base dei metodi di combinazione

Prima di presentare i diversi metodi di combinazione dei *p-value*, è opportuno evidenziare le assunzioni alla base di tali procedure.

Un presupposto fondamentale dei metodi di combinazione dei *p-value* è l'indipendenza tra di essi. La presenza di correlazione, o più in generale di dipendenza, tra i *p-value* potrebbe infatti compromettere la validità teorica e le proprietà inferenziali dei metodi di combinazione che verranno illustrati nel seguito.

L'assunto di indipendenza implica che i singoli *p-value* non siano tra loro collegati, ossia che non vi sia alcuna forma di correlazione o dipendenza tra essi. In altri termini, sotto l'ipotesi nulla globale, la distribuzione di ciascun *p-value* non deve essere influenzata dai valori assunti dagli altri.

Si richiama quindi il concetto di indipendenza statistica di due (o più) variabili aleatorie [Feller (1991),pp. 85-88]. Applicato al caso dei *p-value* sotto l'ipotesi nulla globale, esso può essere espresso come

$$\mathbb{P}(p_1 \leq t_1, \ldots, p_m \leq t_m) = \prod_{i=1}^{m} \mathbb{P}(p_i \leq t_i),$$

per ogni scelta di $t_1, \ldots, t_m$ nell'intervallo $[0, 1]$.

Si intende ora mostrare che, sotto l'ipotesi nulla, la distribuzione dei *p-value* è una Uniforme $(0, 1)$. A tal fine si ricorre al concetto di funzione di ripartizione, attraverso il quale è possibile determinare la probabilità cumulata associata a una variabile casuale X [Deisenroth et al. (2020), pp. 181-182]. Nel presente contesto l'attenzione sarà rivolta esclusivamente al caso di variabili continue.

La funzione di ripartizione di una variabile casuale X, indicata con $F_X(x)$, è definita come [Montgomery & Runger (2010); Keele (2015)]:

$$F_X(x) = \mathbb{P}(X \leq x)\,.$$

Per definizione, la funzione di ripartizione soddisfa le seguenti proprietà fondamentali:

(1) $\lim_{x \to -\infty} F_X(x) = 0 \quad \text{e} \quad \lim_{x \to +\infty} F_X(x) = 1\,;$

(2) $0 \leq F_X(x) \leq 1\,;$

(3) se $x \leq y$, allora $F_X(x) \leq F_X(y)\,.$

Passando alla dimostrazione della distribuzione del *p-value*, si mostra che esso segue una distribuzione uniforme sotto l'ipotesi nulla, considerando il caso di un'alternativa unilaterale destra.

Teorema 2.1. *Sia $F_T(t) = \mathbb{P}(T \leq t \mid H_0)$ la funzione di ripartizione della statistica test T sotto l'ipotesi nulla H_0, assunta vera. Si supponga che F_T sia continua e strettamente crescente, così da ammettere inversa F_T^{-1}.*

Allora la variabile aleatoria definita come

$$P = 1 - F_T(T)$$

soddisfa

$$P \overset{H_0}{\sim} \text{Uniforme}(0,1)\,.$$

Dimostrazione. Sotto H_0, per ogni $0 \leq u \leq 1$ si ha

$$\mathbb{P}\big(F_T(T) \leq u \mid H_0\big) = \mathbb{P}\big(T \leq F_T^{-1}(u) \mid H_0\big)\,.$$

Per definizione di funzione di ripartizione,

$$\mathbb{P}\big(T \leq F_T^{-1}(u) \mid H_0\big) = F_T\big(F_T^{-1}(u)\big) = u\,.$$

Ne segue che

$$F_T(T) \overset{H_0}{\sim} \text{Uniforme}(0,1)\,.$$

Poiché $P = 1 - F_T(T)$, per ogni $0 \leq p \leq 1$ risulta

$$F_P(p) = \mathbb{P}(P \leq p \mid H_0) = \mathbb{P}\big(1 - F_T(T) \leq p \mid H_0\big) = \mathbb{P}\big(F_T(T) \geq 1 - p \mid H_0\big)\,.$$

Pertanto,

$$\mathbb{P}\big(F_T(T) \geq 1 - p \mid H_0\big) = 1 - \mathbb{P}\big(F_T(T) \leq 1 - p \mid H_0\big) = 1 - (1 - p) = p\,.$$

Si conclude quindi che

$$F_P(p) = p, \qquad 0 \leq p \leq 1\,,$$

ossia

$$P \overset{H_0}{\sim} \text{Uniforme}(0,1)\,.$$

□

2.2 Metodo di Bonferroni

Il metodo di Bonferroni [Bonferroni (1936)] può essere interpretato come un metodo di combinazione dei *p-value* volto a testare l'ipotesi nulla globale

$$H_0 = \bigcap_{i=1}^{m} H_{0,i}\,.$$

L'idea alla base del metodo è estremamente semplice: l'evidenza contro l'ipotesi globale viene valutata considerando il più piccolo tra i *p-value* individuali.

Intuitivamente, se almeno uno dei *p-value* assume un valore sufficientemente piccolo, ciò costituisce evidenza contro l'ipotesi che tutte le ipotesi nulle individuali siano vere. Il metodo di Bonferroni formalizza tale idea utilizzando la disuguaglianza di Boole [Boole (1847)].

Si definisce quindi la statistica di combinazione

$$T_{Bon} = \min_{1 \le i \le m} p_i \,,$$

e si rifiuta l'ipotesi nulla globale quando

$$T_{Bon} \le \frac{\alpha}{m} \,.$$

Il risultato seguente mostra la validità statistica del metodo come test per l'ipotesi globale.

Teorema 2.2. *Siano $p_1, \ldots, p_m$ p-value validi, tali che*

$$\mathbb{P}_{H_0}(p_i \le t) \le t \quad \textit{per ogni } t \in [0,1] \,.$$

Allora il test che rifiuta H_0 se

$$\min_i p_i \le \frac{\alpha}{m}$$

è un test di livello α per l'ipotesi nulla globale, ossia

$$\mathbb{P}_{H_0}\Big(\min_i p_i \le \frac{\alpha}{m}\Big) \le \alpha \,.$$

Dimostrazione. Si osservi che

$$\Big\{\min_i p_i \le \frac{\alpha}{m}\Big\} = \bigcup_{i=1}^{m} \Big\{p_i \le \frac{\alpha}{m}\Big\} \,.$$

Pertanto,

$$\mathbb{P}_{H_0}\Big(\min_i p_i \le \frac{\alpha}{m}\Big) = \mathbb{P}_{H_0}\left(\bigcup_{i=1}^{m} \Big\{p_i \le \frac{\alpha}{m}\Big\}\right) \,.$$

Per la disuguaglianza di Boole,

$$\mathbb{P}_{H_0}\left(\bigcup_{i=1}^{m} \Big\{p_i \le \frac{\alpha}{m}\Big\}\right) \le \sum_{i=1}^{m} \mathbb{P}_{H_0}\Big(p_i \le \frac{\alpha}{m}\Big) \,.$$

Poiché ciascun p_i è valido,

$$\mathbb{P}_{H_0}\Big(p_i \leq \frac{\alpha}{m}\Big) \leq \frac{\alpha}{m}\,.$$

Ne segue

$$\sum_{i=1}^{m} \frac{\alpha}{m} = \alpha\,,$$

da cui la tesi. □

Un modo alternativo di formulare il metodo consiste nel definire il *p-value combinato*

$$p_{Bon} = m \cdot \min_i p_i\,.$$

Poiché p_{Bon} può eccedere 1, si considera

$$p_{Bon} = \min\{1,\, m \min_i p_i\}\,.$$

Il rifiuto dell'ipotesi nulla globale avviene quindi quando

$$p_{Bon} \leq \alpha\,.$$

Se inoltre i *p-value* sono indipendenti e

$$p_i \overset{H_0}{\sim} \text{Uniforme}(0,1)\,,$$

allora

$$\mathbb{P}_{H_0}\Big(\min_i p_i \leq t\Big) = 1 - (1-t)^m\,.$$

Ponendo $t = \alpha/m$ si ottiene

$$\mathbb{P}_{H_0}\Big(\min_i p_i \leq \frac{\alpha}{m}\Big) = 1 - \Big(1 - \frac{\alpha}{m}\Big)^m\,.$$

Per valori piccoli di α,

$$1 - \Big(1 - \frac{\alpha}{m}\Big)^m \approx \alpha\,,$$

mostrando che il metodo è leggermente conservativo ma asintoticamente accurato per α piccolo.

Il metodo di Bonferroni presenta il vantaggio di non richiedere l'assunzione di indipendenza tra i *p-value* per garantire la validità del test globale. Tuttavia, esso utilizza esclusivamente il più piccolo tra i *p-value*, ignorando l'informazione contenuta negli altri, e può risultare conservativo quando il numero di ipotesi m è elevato.

2.3 Metodo di Šidák

Un metodo alternativo di combinazione dei *p-value*, strettamente collegato a quello di Bonferroni ma meno conservativo sotto indipendenza, è il metodo di Šidák [Šidák (1967)].

Anche in questo caso l'idea consiste nel basare il test globale sul più piccolo tra i *p-value* individuali. Tuttavia, a differenza di Bonferroni, la soglia critica viene determinata sfruttando esplicitamente l'ipotesi di indipendenza tra i *p-value*.

Si definisce dunque la statistica di combinazione

$$T_{Sid} = \min_{1 \le i \le m} p_i \,,$$

e si rifiuta l'ipotesi nulla globale

$$H_0 = \bigcap_{i=1}^{m} H_{0,i}$$

quando

$$T_{Sid} \le \alpha_{Sid} \,,$$

dove la soglia di Šidák è data da

$$\alpha_{Sid} = 1 - (1 - \alpha)^{1/m} \,.$$

Sotto l'ipotesi nulla globale e assumendo indipendenza,

$$p_1, \ldots, p_m \overset{\text{i.i.d.}}{\sim} \text{Uniforme}(0, 1) \,.$$

In tal caso, la distribuzione del minimo è nota:

$$\min(p_1, \ldots, p_m) \sim \text{Beta}(1, m) \,.$$

Infatti,

$$\mathbb{P}_{H_0}\Big(\min_i p_i \le t\Big) = 1 - (1 - t)^m \,, \qquad 0 \le t \le 1 \,.$$

Teorema 2.3. *Sotto l'ipotesi nulla globale e assumendo indipendenza dei p-value, il test che rifiuta H_0 se*

$$\min_i p_i \le 1 - (1 - \alpha)^{1/m}$$

è un test di livello esattamente α, ossia

$$\mathbb{P}_{H_0}\Big(\min_i p_i \leq 1-(1-\alpha)^{1/m}\Big) = \alpha\,.$$

Dimostrazione. Si ha

$$\mathbb{P}_{H_0}\Big(\min_i p_i \leq t\Big) = 1 - \mathbb{P}_{H_0}\left(\bigcap_{i=1}^{m}\{p_i > t\}\right)\,.$$

Per indipendenza,

$$= 1 - \prod_{i=1}^{m}\mathbb{P}_{H_0}(p_i > t)\,.$$

Poiché $p_i \sim \text{Uniforme}(0,1)$,

$$\mathbb{P}_{H_0}(p_i > t) = 1 - t\,.$$

Dunque

$$\mathbb{P}_{H_0}\Big(\min_i p_i \leq t\Big) = 1-(1-t)^m\,.$$

Ponendo

$$t = 1-(1-\alpha)^{1/m}\,,$$

si ottiene

$$\mathbb{P}_{H_0}\Big(\min_i p_i \leq 1-(1-\alpha)^{1/m}\Big) = 1 - \left[(1-\alpha)^{1/m}\right]^m = 1-(1-\alpha) = \alpha\,.$$

□

Analogamente al caso di Bonferroni, si può definire un *p-value combinato* di Šidák come

$$p_{Sid} = 1-(1-\min_i p_i)^m\,.$$

Il rifiuto dell'ipotesi nulla globale avviene quindi quando

$$p_{Sid} \leq \alpha\,.$$

Il metodo di Šidák risulta meno conservativo rispetto a Bonferroni, poiché

$$1-(1-\alpha)^{1/m} \geq \frac{\alpha}{m}\,.$$

Tale vantaggio deriva dall'utilizzo dell'ipotesi di indipendenza, che consente di ottenere un controllo esatto del livello del test. In assenza di indipendenza, tuttavia, la validità del metodo non è più garantita, mentre il metodo di Bonferroni rimane valido.

2.4 Metodo di Simes

Dopo aver illustrato i metodi di Bonferroni e di Šidák,entrambi basati sul confronto tra il minimo dei *p-value* e una soglia opportunamente corretta,è naturale considerare procedure alternative che utilizzino in modo più articolato l'informazione contenuta nell'insieme dei *p-value*.

In questa prospettiva si colloca il metodo proposto da Simes [Simes (1986)], che può essere interpretato come un'estensione dell'approccio fondato sul minimo dei *p-value*. A differenza di Bonferroni e Šidák, il metodo di Simes introduce soglie dipendenti dal rango dei *p-value* ordinati, consentendo una valutazione globale dell'evidenza contro l'ipotesi nulla attraverso un unico criterio decisionale.

Sia

$$H_0 = \bigcap_{i=1}^{m} H_{0,i}$$

l'ipotesi nulla globale e siano $p_1, \ldots, p_m$ i *p-value* associati ai singoli test. Si considerino i *p-value* ordinati in senso crescente:

$$p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(m)} \, .$$

Il metodo di Simes rifiuta l'ipotesi nulla globale quando esiste un indice i tale che

$$p_{(i)} \leq \frac{i}{m}\alpha \, .$$

Il metodo può essere espresso in forma equivalente attraverso la statistica di combinazione

$$T_{Simes} = \min_{1 \leq i \leq m} \left(\frac{m}{i} p_{(i)} \right) \, .$$

Il *p-value globale* associato al metodo di Simes è quindi

$$p_{Simes} = T_{Simes} \, ,$$

e il rifiuto dell'ipotesi nulla globale avviene quando

$$p_{Simes} \leq \alpha \, .$$

Teorema 2.4. *Sotto l'ipotesi nulla globale, se*

$$p_1, \ldots, p_m \overset{i.i.d.}{\sim} \text{Uniforme}(0,1) \, ,$$

allora la statistica

$$T_{Simes} = \min_{1 \leq i \leq m} \left(\frac{m}{i} p_{(i)} \right)$$

soddisfa

$$\mathbb{P}_{H_0}(T_{Simes} \leq \alpha) \leq \alpha \, ,$$

ossia il test di Simes è di livello α.

La dimostrazione si basa sulle proprietà congiunte degli ordini statistici di variabili Uniformi indipendenti e sulla disuguaglianza di Simes [Simes (1986); Sarkar (2008)].

Il metodo di Simes risulta meno conservativo rispetto a Bonferroni, poiché la soglia critica cresce con l'indice i. In particolare, mentre Bonferroni utilizza sempre la soglia α/m, Simes adotta soglie progressive $\frac{i}{m}\alpha$, che permettono di valorizzare l'informazione contenuta nei *p-value* di rango superiore.

Tale caratteristica rende il metodo più potente in presenza di più ipotesi debolmente significative, pur mantenendo il controllo del livello del test sotto indipendenza (o sotto particolari forme di dipendenza positiva).

2.5 Metodo di Fisher

Dopo aver esaminato metodi fondati sul confronto tra il minimo dei *p-value* e soglie opportunamente definite,come nel caso di Bonferroni e Šidák,si consideri ora un approccio concettualmente differente.

Il metodo proposto da Fisher [Fisher (1934)] si discosta infatti dalle procedure basate esclusivamente sul *p-value* più piccolo, pur condividendo con il metodo di Simes l'idea di utilizzare l'informazione contenuta in tutti i *p-value*. In particolare, l'evidenza complessiva contro l'ipotesi nulla globale H_0 viene sintetizzata combinando tutti i *p-value* attraverso una trasformazione logaritmica, che conduce a una statistica test con distribuzione nota sotto H_0.

L'idea fondamentale si basa sul risultato secondo cui la somma di variabili casuali indipendenti con distribuzione χ^2 è ancora distribuita come una variabile χ^2, con un numero di gradi di libertà pari alla somma dei gradi di libertà individuali [Fisher (1934)].

Siano $p_1, \ldots, p_m$ i *p-value* associati ai singoli test. Il metodo di Fisher definisce la statistica di combinazione

$$T = -2 \sum_{i=1}^{m} \ln p_i \,.$$

Sotto l'ipotesi nulla globale e assumendo indipendenza tra i *p-value*, la statistica T viene confrontata con la distribuzione χ^2 con $2m$ gradi di libertà.

Il test rifiuta l'ipotesi nulla globale quando

$$T > \chi^2_{2m}(1-\alpha)\,,$$

dove $\chi^2_{2m}(1-\alpha)$ rappresenta il quantile di ordine $1-\alpha$ della distribuzione χ^2 con $2m$ gradi di libertà.

Teorema 2.5. *Siano*

$$p_1, \ldots, p_m \overset{i.i.d.}{\sim} \text{Uniforme}(0,1) \quad \textit{sotto } H_0\,.$$

Allora la statistica

$$T = -2 \sum_{i=1}^{m} \ln p_i$$

segue, sotto H_0, una distribuzione

$$T \sim \chi^2_{2m}\,.$$

Dimostrazione. Sotto l'ipotesi nulla globale,

$$p_i \overset{\text{i.i.d.}}{\sim} \text{Uniforme}(0,1)\,.$$

È noto che, se $U \sim \text{Uniforme}(0,1)$, allora

$$-\ln U \sim \text{Exp}(1)\,,$$

dove Exp(1) indica la distribuzione esponenziale con parametro 1.

Inoltre, una variabile casuale con distribuzione χ^2_2 è equivalente a una variabile esponenziale con parametro $1/2$. Ne segue che

$$-2\ln p_i \sim \chi^2_2 \,.$$

Poiché i p_i sono indipendenti, anche le variabili $-2\ln p_i$ sono indipendenti. Essendo T la somma di m variabili indipendenti χ^2_2, si ottiene

$$T \sim \chi^2_{2m} \,.$$

□

Il metodo di Fisher può essere espresso anche in termini di *p-value combinato*.

Definita la statistica osservata T_{obs}, il *p-value globale* è dato da

$$p_{Fisher} = \mathbb{P}\big(\chi^2_{2m} \geq T_{obs}\big) \ .$$

Il rifiuto dell'ipotesi nulla globale avviene quindi quando

$$p_{Fisher} \leq \alpha \,.$$

Il metodo di Fisher utilizza l'informazione contenuta in tutti i *p-value* e tende a essere particolarmente efficace quando più test mostrano evidenza moderata contro l'ipotesi nulla.

Tuttavia, a differenza del metodo di Bonferroni, esso richiede l'assunzione di indipendenza tra i *p-value* per garantire la validità esatta della distribuzione della statistica test.

2.6 Metodo di Stouffer

Un ulteriore metodo di combinazione dei *p-value* per testare l'ipotesi nulla globale è quello proposto da Stouffer et al. [Stouffer et al. (1949)].

Analogamente al metodo di Fisher, l'approccio di Stouffer utilizza l'informazione contenuta in tutti i *p-value*, ma invece di applicare una trasformazione logaritmica, li trasforma in statistiche z attraverso la funzione quantile della distribuzione normale standard.

Siano $p_1, \ldots, p_m$ i *p-value* associati ai singoli test. La statistica di combinazione di Stouffer è definita come

$$Z_{Sto} = \frac{1}{\sqrt{m}} \sum_{i=1}^{m} \Phi^{-1}(1 - p_i)\,,$$

dove Φ^{-1} indica la funzione quantile della distribuzione normale standard.

Assumendo un'alternativa unilaterale destra, il test rifiuta l'ipotesi nulla globale

$$H_0 = \bigcap_{i=1}^{m} H_{0,i}$$

quando

$$Z_{Sto} > z_{1-\alpha}\,,$$

dove

$$z_{1-\alpha} = \Phi^{-1}(1 - \alpha)\,.$$

Teorema 2.6. *Siano*

$$p_1, \ldots, p_m \overset{i.i.d.}{\sim} \text{Uniforme}(0,1) \quad \textit{sotto } H_0\,.$$

Allora la statistica combinata

$$Z = \frac{1}{\sqrt{m}} \sum_{i=1}^{m} \Phi^{-1}(1 - p_i)$$

segue, sotto H_0, una distribuzione normale standard:

$$Z \sim \text{Normale}(0,1)\,.$$

Dimostrazione. Sotto l'ipotesi nulla globale,

$$p_i \overset{\text{i.i.d.}}{\sim} \text{Uniforme}(0,1)\,.$$

Definiamo

$$Z_i = \Phi^{-1}(1 - p_i)\,.$$

Poiché $1 - p_i$ è ancora distribuito come Uniforme$(0,1)$ e la funzione quantile trasforma una variabile uniforme in una variabile normale standard, si ha

$$Z_i \sim \text{Normale}(0,1)\,,$$

con indipendenza tra le variabili Z_i.

La somma di m variabili indipendenti Normale$(0, 1)$ è distribuita come

$$\sum_{i=1}^{m} Z_i \sim \text{Normale}(0, m)\,.$$

Dividendo per $\sqrt{m}$ si ottiene

$$Z = \frac{1}{\sqrt{m}} \sum_{i=1}^{m} Z_i \sim \text{Normale}(0, 1)\,.$$

□

Dato il valore osservato Z_{obs}, il *p-value globale* associato al metodo di Stouffer è

$$p_{Sto} = 1 - \Phi(Z_{obs})\,,$$

nel caso di alternativa unilaterale destra.

L'ipotesi nulla globale viene rifiutata quando

$$p_{Sto} \leq \alpha\,.$$

Il metodo di Stouffer richiede l'assunzione di indipendenza tra i *p-value* per garantire la validità esatta della distribuzione della statistica combinata.

Rispetto al metodo di Fisher, esso assegna un peso lineare alle evidenze individuali e può risultare più sensibile quando i segnali sono deboli ma coerenti nella stessa direzione.

Capitolo 3

Disegno dello studio di simulazione

Nel presente capitolo si illustra quali sono le caratteristiche dello studio di simulazione che ha come obiettivo quello di rispondere alla domanda statistica operativa. Per cui, dopo aver illustrato quest'ultima, si individueranno le ipotesi nulla ed alternativa, i parametri utilizzati nella simulazione, il modello di generazione dei dati, il modo con cui vengono ottenuti i *p-value* ed, infine, l'applicazione dei metodi di combinazione di *p-value* indipendenti. Nello specifico, per lo studio di simulazione si terrà conto di tre casistiche a seconda del valore del parametro m e di come viene determinato il valore di θ_i.

$\Diamond$ $m = 2$

Definizione 3.1. Si considerino X_1, X_2 v.a. indipendenti con distribuzione $N(\theta_i, 1)$ *per* $i = 1, 2$. Sia $\theta = (\theta_1, \theta_2)$ il vettore delle medie.

La media del vettore è sparsa, tale che $\|\theta\|_0 = k$ con un livello di *sparsity* $k \in \{1, 2\}$. La forza complessiva del segnale è controllata da $\|\theta\|_1 = h$, dove $h > 0$. Specificamente:

- Se $k = 1$, una sola componente di θ è diversa da zero ed è pari a h.
- Se $k = 2$, entrambe le componenti di θ sono diverse da zero e, quindi, soddisfano la condizione che $|\theta_1| + |\theta_2| = h$.

$\Diamond$ $m > 2$**, omogenea**

Definizione 3.2. Si considerino m v.a. indipendenti $X_1, \ldots, X_m$ dove X_i ha distribuzione $N(\theta_i, 1)$ *per* $i = 1, \ldots, m$. Sia $\theta = (\theta_1, \ldots, \theta_m)$ il vettore delle medie.

La media del vettore è sparsa, tale che $\|\theta\|_0 = k$ con un livello di *sparsity* $k \in [m]$. La forza complessiva del segnale è controllata da $\|\theta\|_1 = h$, dove $h > 0$.

Specificamente, il segnale è distribuito uniformemente lungo i componenti non nulli: $\theta_i = h/k$ per $i = 1, \ldots, k$ e 0 per $i = k+1, \ldots, m$.

$\Diamond$ $m > 2$, lineare

Definizione 3.3. Si considerino m v.a. indipendenti $X_1, \ldots, X_m$ dove X_i ha distribuzione $N(\theta_i, 1)$ *per* $i = 1, \ldots, m$. Sia $\theta = (\theta_1, \ldots, \theta_m)$ il vettore delle medie.

La media del vettore è sparsa, tale che $\|\theta\|_0 = k$ con un livello di *sparsity* $k \in [m]$. La forza complessiva del segnale è controllata da $\|\theta\|_1 = h$, dove $h > 0$. Specificamente, il segnale è distribuito in modo lineare crescente proporzionalmente a i lungo i componenti non nulli: $\theta_i = \frac{i \cdot 2h}{k(k+1)}$ per $i = 1, \ldots, k$ e 0 per $i = k+1, \ldots, m$.

Definizione di norma

Nel corso di queste ultime definizioni, si è fatto uso del concetto di *norma* che è stato rappresentato con $\|\theta\|_0$ e $\|\theta\|_1$ le quali sono le norme L_0 e L_1 del vettore delle medie $\theta = (\theta_1, \ldots, \theta_m)$. Per darne una definizione generale si utilizzerà x al posto di θ.

Una *norma* è una funzione definita su uno spazio vettoriale a valori reali non negativi che soddisfa le proprietà di positività, omogeneità e disuguaglianza triangolare. Per cui, gli spazi L_p sono spazi vettoriali di funzioni misurabili la cui p-esima potenza è integrabile. Quindi,la norma L_p è una generalizzazione naturale delle *p-norm* definite su spazi vettoriali di dimensioni finite.[Riesz (1910);Dunford & Schwartz (1988)].

Nello specifico, sia $\mathbf{x} = (x_1, \ldots, x_m) \in \mathbb{R}^m$, con $m \in \mathbb{N}, m \geq 1$, allora la *p-norm*, ossia la norma L_p, dove $p \geq 1$, è definita da

$$\|x\|_p = \left(\sum_{i=1}^{n} |x_i|^p \right)^{1/p} .$$

Limitando la definizione generale ai casi di L_0 e L_1 è possibile affermare che:

- $p = 0$, essendo $p \geq 1$ una delle condizioni della definizione di *p-norm*, questo caso rappresenta un'eccezione se non un errore vero e proprio. Per cui, a L_0 viene dato un significato più figurato di *norma*, ossia il fatto che essa calcola il numero di elementi non nulli[Donoho (2006)]. Ottenendo quindi che

$$\|x\|_0 = \textit{conteggio di } x_i \neq 0 \, .$$

Tale concetto verrà ripreso nel corso del capitolo quando si approfondirà la c.d. *sparsity*.

- $p = 1$, attraverso tale uguaglianza, si semplifica la formula generale della *p-norm* in una semplice sommatoria degli elementi del vettore, che rappresenta anche l'effetto totale del vettore:

$$\|x\|_1 = \sum_{i=1}^{n} |x_i| = h\,.$$

3.1 Domanda statistica operativa

In sintesi, si potrebbe affermare che la domanda statistica operativa a cui si desidera rispondere tramite lo studio di simulazione è:

Tra i metodi di combinazione dei *p-value* indipendenti di Bonferroni, Sidak, Fisher, Stouffer o Simes. Quale, e in quali condizioni (rispetto a numero di test, *sparsity* e dimensione dell'effetto totale h), presenta la maggiore potenza del test ad ipotesi nulla globale?

La domanda statistica quindi ha l'obiettivo di individuare come cambia l'"ottimalità" dei diversi metodi a seconda dei parametri di *input* della simulazione, ossia quale metodo combinatorio garantisce una potenza del test maggiore rispetto alle altre considerando un'ipotesi nulla globale, di cui si dirà nel prossimo paragrafo.

3.2 Ipotesi nulla ed alternativa

Il test d'ipotesi alla base dello studio di simulazione sarà un test ad alternativa destra (anche se potrebbe essere estendibile agli altri casi) su un parametro θ, il cui valore cambierà a seconda dei diversi test. Per cui si potrebbe affermare che, avendo m test, questi singolarmente assumono la seguente forma:

$$\begin{cases} H_{0,i} : \theta_i \leq \theta_{0,i} \\ H_{1,i} : \theta_i > \theta_{1,i} \end{cases} \quad con\ i = 1, \ldots, m$$

Dato che, come si è detto sopra, la simulazione raccoglierà i singoli test sotto un'unica ipotesi nulla globale, ovvero che tutte le ipotesi nulle siano vere contro l'ipotesi alternativa che almeno un'ipotesi nulla sia falsa, è possibile rappresentare il test su cui effettivamente si baserà lo studio di simulazione sui diversi metodi

combinatori come:

$$\begin{cases} H_0^{globale} : H_{01} \cap H_{02} \cap \cdots \cap H_{0m} \\ H_1^{globale} \; \exists i \; tale \; che \; H_{0i} \; è \; falsa \end{cases} \qquad con \; i = 1, \ldots, m$$

Quindi ricapitolando, l'ipotesi nulla che viene utilizzata nello studio di simulazione è l'ipotesi nulla globale per cui nessuna ipotesi nulla è falsa, mentre l'ipotesi alternativa postula che esista almeno un caso i in cui l'ipotesi nulla non sia vera.

3.3 Parametri della simulazione

Prima di mostrare il modello di generazione dei dati e il conseguente ottenimento dei *p-value*, si procede a illustrare brevemente le caratteristiche dei parametri selezionati per lo studio di simulazione.

- m, rappresenta il numero di test svolti per verificare l'ipotesi nulla globale

- α, come si è più volte detto finora, è il livello di significatività di riferimento per il test d'ipotesi. Nello specifico, durante la simulazione sarà fissato ad un valore di 0.1.

- h, ossia la dimensione dell'effetto che viene rappresentato dal valore che possono assumere in totale i diversi θ_i. Per cui con $m = 2$ la dimensione dell'effetto sarà fissa (nella simulazione) e pari a $\theta_1 + \theta_2$. Tale parametro è da intendersi equivalente a quanto illustrato all'inizio del presente capitolo, in termini di "effetto del vettore" e di L_1.

- B, rappresenta il numero di repliche per singola simulazione. Come si vedrà, i diversi test d'ipotesi verranno ripetuti per $B = 10^4$ volte, in modo da avere un'attendibilità statistica dei risultati. Infatti, tale valore permette di stimare un'approssimazione dei risultati molto ridotta attraverso il calcolo dell'errore Monte Carlo [per approfondire l'argomento, vedere Geyer (1992)] che è l'incertezza dovuta all'uso di un numero finito di simulazioni, calcolabile come $\sqrt{B^{-1} \cdot \hat{Pow}(1 - \hat{Pow})}$ [Koehler et al. (2009)]. Nello specifico, con 10000 repliche e una stima di potenza pari a 0.5, l'errore standard Monte Carlo sarà 0.005. Pertanto, l'oscillazione casuale della stima (nel caso "peggiore" possibile, cioè $\hat{Pow} = 0.5$) è dell'ordine di ± 0.005 (deviazione standard) e pari a ± 0.01 ($2 \cdot SE_{MC}$) considerando un intervallo di confidenza approssimativo al 95%.

Questa precisione è considerata sufficientemente affidabile per lo studio qui svolto.

- *Potenza*, la definizione operativa di potenza in questo contesto è: individuare in media quante volte il metodo combinatorio utilizzato sia in grado di rifiutare l'ipotesi nulla globale, che in termini relativi (rispetto a 1) indica la probabilità che il metodo analizzato sia in grado di identificare un effetto reale (rifiutando H_0).

Anche se non rappresenta propriamente un parametro di "input" della simulazione, appare necessario riprendere il concetto di *sparsity*, anche col fine di comprendere meglio le varie casistiche/simulazioni svolte.

La c.d. *sparsity*, in questo contesto, potrebbe considerarsi come la densità dell'effetto, infatti tale valore rappresenta quanti dei parametri del modello (ergo simulazione) hanno un effetto/valore pari a zero [Gale et al. (2019); per una definizione matematica vedi McCullagh & Polson (2018)]. Per cui, tanto maggiore sarà la *sparsity*, tanto minori saranno i parametri con un effetto, concludendo che la distribuzione dei valori sarà molto concentrata su pochi parametri (θ_i). Ad esempio, con $m = 2$ la sparsity sarà al suo valore massimo se $\theta_1 = h$ e $\theta_2 = 0$ o viceversa.

Il complementare della *sparsity* in termini di conteggio dei parametri non nulli verrà rappresentato da k, su cui si focalizzerà la simulazione. Per fare un esempio, con $m = 10$ e solo $\theta_1, \theta_3, \theta_5$ sono nulli, allora $k = m - 3 = 7$.

Dopo aver individuato gli elementi caratteristici della simulazione, appare logico procedere con la descrizione di quest'ultima. Nello specifico, come già illustrato formalmente ad inizio capitolo, lo studio si focalizza su due scenari, di cui uno si divide a sua volta in altri due, ossia:

- $m = 2$, in questo caso, essendo il numero di test svolti pari a 2, anche i parametri θ_i saranno due e la somma dei loro valori sarà h, per cui $\theta_1 = h - \theta_2$ e viceversa. Quindi si ottiene una simulazione che può focalizzarsi su uno solo dei due parametri per lo studio, in quanto l'altro parametro sarà speculare e perciò sostanzialmente identico.

- $m > 2$, in questo scenario verranno considerati più di 2 test per simulazione. Quindi per agevolare il trattamento dei parametri (che saranno sicuramente più di 2) è necessario svolgere delle semplificazioni, ossia scegliere se θ_i sarà:

- *omogenea*, attraverso tale semplificazione si desidera porre che ciascun $\theta_i \neq 0$ assuma un valore, per l'appunto, omogeneo pari a h/k, con k parametri non nulli.
- *lineare*, con questa approssimazione viene, invece, posto un valore "linearmente" crescente. Con k parametri non nulli, il valore di h sarà distribuito linearmente in funzione di i, per cui il valore dei singoli parametri sarà $\theta_i = \frac{i \cdot 2h}{k(k+1)}$.

3.4 Modello di generazione dei dati e dei rispettivi *p-value*

Per svolgere le varie simulazioni bisognerà preliminarmente inserire degli "oggetti" all'interno della funzione di R affinché il programma possa eseguire correttamente il ciclo *for*. Per la simulazione con $m = 2$ avendo come *input* anche h, B, α, sarà sufficiente inserire una griglia relativa ai valori che θ_1 può assumere, attraverso una sequenza da 0 a h con un "passo" tra valori consecutivi di 0.1. Allo stesso modo andrà creata la sequenza relativa al numero di cicli che si desiderano svolgere ossia B, questa volta con un incremento di valore di 1. Dopodiché andranno inserite due (come il numero di test) matrici per i valori dei *p-value* con il numero di righe e colonne esattamente pari alla lunghezza della griglia e dei cicli svolti. Infine, andranno create delle matrici per i diversi metodi combinatori, con le medesime dimensioni delle matrici dei *p-value*. Il ciclo su cui si baserà la simulazione sarà in funzione di i lungo la griglia dei valori di θ_i e di j lungo la sequenza dei cicli. Ottenendo, quindi, una ripetizione di ogni valore della griglia, svolta per B volte.

Mentre, per la simulazione con $m > 2$ sarà sufficiente inserire solamente una sequenza che indichi quanti k test non nulli ci sono $[1 : m]$ e le matrici dei vari metodi di combinazione con righe pari alla lunghezza di k (numero intero) e colonne pari ad un totale di B. In tal caso, il ciclo *for* si basera su i test non nulli, ossia i diversi valori di k lungo la sequenza, e su j numero di cicli. Quindi, riassumendo, una ripetizione per ogni valore di k svolta per B volte.

3.4.1 Modello di generazione dei dati

Il modello di riferimento per la generazione dei dati è "*rnorm*", il quale dà come risultato dei dati distribuiti normalmente con valori di *default* di una standard $N(0, 1)$. Tali valori vengono utilizzati nella generazione dei dati nei casi di $m > 2$ dove verrà generato un vettore di dati di lunghezza m con distribuzione standard.

Mentre nel caso di $m = 2$, data la minore complessità, è possibile generare per ogni test una serie di dati singoli con media pari al valore di θ_1 o di θ_2 a seconda del test che si sta considerando.

Riassumendo quindi, il modello di riferimento sarà:

- *rnorm(m) per* $m > 2$
- *rnorm(1,mean=*θ_i*) per* $m = 2$ *e quindi con* $i = 1, 2$.

3.4.2 Come vengono ottenuti i *p-value*

Per quanto riguarda l'ottenimento dei *p-value* anche in tal caso va svolta una distinzione tra le diverse casistiche. Infatti, con $m = 2$ l'ottenimento è immediato tramite il complementare di *pnorm*[1] che è la funzione di ripartizione delle normali, la quale restituisce il valore dell'area alla sinistra del punto selezionato e, quindi, avendo come riferimento il test ad alternativa destra, il *p-value* sarà pari a $1 - pnorm$.

Nel caso di $m > 2$ bisognerà considerare il caso di più parametri senza effetto, dato che nel caso precedente se uno dei due θ_i fosse stato nullo, l'altro necessariamente sarebbe stato $\neq 0$. Per far fronte a tale problematica, la soluzione proposta è quella di generare i dati come una standard (come visto sopra) ed aggiungere, prima di calcolare il *p-value*, il parametro θ_i in base a quanti di tali parametri siano nulli o meno, cioè in funzione di k. Per cui, nel caso dell'omogenea, essendo $\theta_i = h/k \ \forall i \in \{1, \ldots, k\}$, sarà sufficiente ripetere tale valore per k volte e considerare gli altri paramentri $(m - k)$ pari a zero. Mentre, qualora si considerasse il valore di θ_i come lineare, sarà sufficiente inserire solo θ_i nel vettore dei dati in quanto esso avrà valori diversi per k volte, e sarà ancora necessario inserire $m - k$ volte il valore nullo.

Riprendendo quanto si è detto nella sezione precedente e quella corrente, si può riassumere che i *p-value* vengono ottenuti attraverso:

- $pvalue = 1 - pnorm(rnorm(1, mean = \theta_i))$ *per* $m = 2$ *e quindi con* $i = 1, 2$
- $pvalue = 1 - pnorm(z)$ *per* $m > 2$ *omogenea, dove* $z = rnorm(m) + c(rep(\theta_i, k[i]), rep(0, m - k[i]))$
- $pvalue = 1 - pnorm(z)$ *per* $m > 2$ *lineare, dove* $z = rnorm(m) + c(\theta_i, rep(0, m - k[i]))$.

[1] La funzione di R *pnorm* è diversa dal concetto di *p-norm* di inizio capitolo, la quale invece rappresenta una norma *p-esima* di un vettore

3.5 Applicazione dei metodi combinatori

Una volta generati i dati ed i conseguenti *p-value* è necessario ora applicare quanto illustrato nel Capitolo 2, senza svolgere alcuna distinzione fra le varie casistiche relative al valore di m di cui sopra.

Nello specifico, si andranno ad applicare le regole di rifiuto dei diversi metodi all'interno della simulazione come riassunto di seguito, considerando quanto finora detto:

- Bonferroni, si rifiuta l'ipotesi globale se

$$min(p_i) \leq \alpha/m\,.$$

- Sidak, si rifiuta l'ipotesi globale se

$$\min_{i=1,\dots,m}(p_i) \leq 1-(1-\alpha)^{\frac{1}{m}}\,.$$

- Simes, si rifiuta l'ipotesi globale se

$$\min_{1\leq i\leq m}\left(\frac{m}{i}p_{(i)}\right) \leq \alpha\,.$$

- Fisher, si rifiuta l'ipotesi globale se

$$-2\sum_{i=1}^{m}\ln p_i > \chi^2_{2m}(1-\alpha)\,.$$

- Stouffer, si rifiuta l'ipotesi globale se

$$\frac{1}{\sqrt{m}}\sum_{i=1}^{m} z_i > \Phi^{-1}(1-\alpha) \quad dove\ z_i = \Phi^{-1}(1-p_i)\,.$$

Inoltre, col fine di confrontare la funzione di potenza dei diversi metodi viene incluso il test c.d. "ottimo"[Bittman et al. (2009)], che dipende da parametri definiti arbitrariamente e quindi non applicabile nella pratica. Per tale motivo, il suo ruolo è unicamente quello di *benchmark* teorico, utile a interpretare e contestualizzare le prestazioni degli altri test, e non di un'effettiva alternativa operativa. Il test "ottimo" si basa sul Lemma Neyman-Pearson [Neyman & Pearson (1933)] e prevede di rifiutare

l'ipotesi nulla globale se, considerando $l = 1, \ldots, k_i$,

$$\frac{1}{\sqrt{k_i}} \sum_{l=1}^{k_i} \Phi^{-1}(1 - p_l) > z_{1-\alpha} \, .$$

Come risultato del ciclo *for* si dovrebbero ottenere le matrici dei diversi test, di inizio comando, compilate con dei risultati binari (TRUE/FALSE), ossia se l'ipotesi nulla è stata rifiutata (TRUE) in corrispondenza delle coordinate $[i, j]$. Per cui, dato che R interpreta $TRUE = 1$ e $FALSE = 0$ per calcolare la potenza dei diversi metodi basterà fare la media per ogni riga. In questo modo, si ottiene un vettore che evidenzia come varia la potenza del metodo analizzato in base a i, che può essere la griglia dei valori di θ_i (nel caso di $m = 2$) o la c.d. *sparsity*, ovvero quanti test sono nulli (nel caso di $m > 2$).

Infine, per facilitare la visualizzazione dei risultati si procederà con la costruzione di un *plot* le cui assi saranno per le ordinate la potenza e per le ascisse la *sparsity*/griglia di θ_i. Inoltre vengono posti come limiti dell'asse delle y il valore massimo e minimo della potenza ottenuta, onde evitare un eccessivo intervallo di visualizzazione.

All'interno del grafico saranno inserite delle linee rappresentative dei diversi metodi di combinazione di *p-value* indipendenti con diversi colori ed una legenda che colleghi al metodo il colore corrispondente.

Capitolo 4

Risultati dello studio di simulazione

Nel presente capitolo si illustreranno i risultati più significativi della simulazione (riportati anche nell'Appendice B), distinguendo i diversi scenari in base al numero di test m. In particolare, si presenteranno i grafici di quanto ottenuto e si proporranno delle osservazioni iniziali che verranno discusse in termini teorici nel capitolo successivo.

Nel corso delle varie simulazioni, si è deciso di porre $\alpha = 0.1$ come parametro fisso, così da poter contare su una maggiore potenza (rispetto ad $\alpha = 0.05$) e, quindi, su una migliore visualizzazione dei risultati. Inoltre, per garantire una certa accuratezza dei risultati, si pone $B = 10^4$, in modo che la singola simulazione venga ripetuta 10000 volte per singolo valore della *sparsity*. Per cui, si può concludere che gli unici parametri modificati nelle diverse simulazioni saranno il numero di test m e, quindi, il numero di test non nulli lungo la sequenza di k, ed il valore dell'effetto totale h.

4.1 Scenario con m=2

Tale scenario è quello che permette un grado di precisione della simulazione maggiore rispetto agli altri. Infatti, come è stato spiegato nel Capitolo 3, dove per i casi con $m > 2$ sono necessarie delle semplificazioni riguardo al valore di θ, in questo caso i singoli valori all'interno del vettore possono idealmente essere considerati in una situazione di continuità. Ossia si può aumentare θ_i per un valore talmente piccolo, tale per cui la funzione che rappresenta la potenza dei diversi metodi è continua grazie ad un aumento marginale dei valori lungo l'asse delle ascisse (θ_i) prossimo allo zero.

Nello specifico, per facilitare la visualizzazione dei risultati e data la poca rilevanza della variazione del parametro h relativo all'effetto, si decide di considerare solo il grafico ottenuto da $h = 2$. Da una prima simulazione includendo tutti i metodi fin

qui illustrati, si evince che, come era prevedibile, Bonferroni, Sidak e Simes sono simili tra loro. Per cui si decide di inserire nella simulazione finale solo quello che garantisce una potenza maggiore per quasi ogni valore di θ_1, ossia Simes.

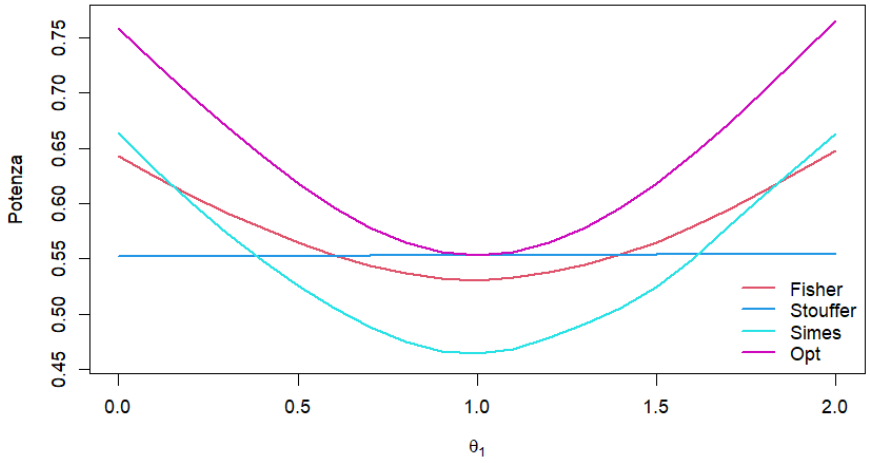

Figura 4.1: Potenza dei metodi con $m = 2$ *e* $h = 2$

Qualora si dividesse il grafico a metà in corrispondenza di $\theta_1 = 1$ si evidenzia come in corrispondenza di tale valore, l'effetto h è equamente distribuito tra i due parametri. Mentre, qualora $\theta_1 = 0$ oppure $\theta_2 = 0$, si ottiene il valore massimo di concentrazione dell'effetto.

Per cui, osservando la prima metà del grafico, si può concludere che il metodo di Simes (così come Bonferroni e Sidak) ha una potenza maggiore in corrispondenza di una forte densità, ossia quando l'effetto è distribuito solo su θ_1 o solo su θ_2. Nel caso di equa ripartizione di h il metodo più potente è Stouffer; mentre quando la *sparsity* è nulla, ma non c'è una tendenziale uguaglianza tra i valori dei due parametri, allora il metodo che ha maggiore probabilità di rifiutare l'ipotesi globale falsa è Fisher. Tali considerazioni non tengono conto del test "ottimo" che ha la potenza maggiore (rispetto agli altri) in ogni condizione, anche se va sottolineato come in corrispondenza di $\theta_1 = \theta_2 = 1$ ha la stessa potenza del metodo di Stouffer.

4.2 Scenario con m>2

La simulazione appena svolta si concentrava solo su due test ed era possibile evidenziare come veniva distribuito h tra i due parametri, andando quindi oltre alle sole due casistiche $k = \{1, 2\}$. Nelle prossime simulazioni, invece, dovendo

considerare $m > 2$ test, si analizzerà come varia la potenza dei metodi in base al valore di k test non nulli, tralasciando quindi come viene distribuito l'effetto h, in quanto quest'ultimo è l'oggetto alla base delle semplificazioni svolte, ossia θ_i omogenea e θ_i lineare.

Nello svolgere le svariate simulazioni, inserendo diversi valori dei parametri *input* (m, h), si nota che, anche in questo caso i metodi "simil-Bonferroni" (Bonferroni, Sidak e Simes) danno risultati molto simili tra loro, per cui anche nello scenario di $m > 2$ con $\theta_i = h/k$ si considererà solamente il metodo (tra i tre appena accennati) che mostra una potenza tendenzialmente maggiore, ossia Simes.

Come prima simulazione si considerano i parametri utilizzati nella precedente, aumentando però la numerosità dei test a $m = 5$.

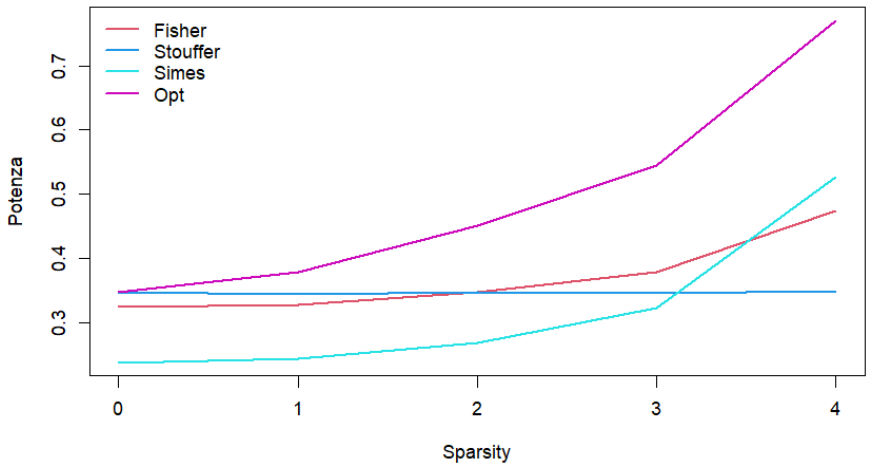

Figura 4.2: Potenza dei metodi con θ_i omogenea, $m = 5$ *e* $h = 2$

Appurato il fatto che il test "ottimo" risulta ovviamente il più potente in ogni condizione. Dai risultati ottenuti, si evidenzia che i metodi combinatori "simil-Bonferroni" sono più potenti nel caso di molti test nulli, ma la loro potenza decresce fortemente al diminuire della *sparsity*. Mentre, il metodo di Fisher, anche se non è il più potente per valori di k prossimi a 1 o a m, è il più forte (dopo il test "ottimo") se $sparsity = 3$ e tende a mantenere una buona potenza all'aumentare del numero di test non nulli. Infine, Stouffer ha un andamento pressoché costante e garantisce la potenza più alta quando $Sparsity < 1$, potenza prossima a quello ottimale.

A questo punto, si prova a porre $m = h$ per comprendere come variano i risultati. Per una migliore visualizzazione di essi, si utilizza $m = h = 5$.

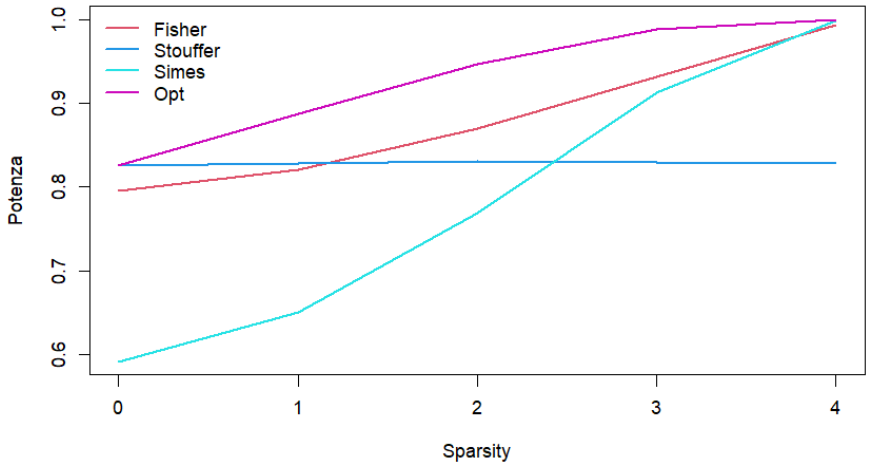

Figura 4.3: Potenza dei metodi con θ_i omogenea e $m = h = 5$

Tralasciando la attesa superiorità del test "ottimo", dal grafico si nota come il metodo di Fisher abbia una potenza superiore per $k < 4$, anche se nel caso di *sparsity* massima condivide tale superiorità coi metodi "simil-Bonferroni", i quali, però, tendono a indebolirsi con l'aumentare di k. Anche in tal caso, Stouffer segue un andamento costante ed è il metodo più potente quando la *sparsity* è al minimo.

Va inoltre sottolineato come siano cambiati i valori della Potenza rispetto al caso precedente. Infatti, all'aumentare del segnale h appare logico che questo incida sulla capacità dei test di individuare almeno un'ipotesi nulla falsa da rifiutare.

Ora, si desidera analizzare cosa accade quando la numerosità dei test viene aumentata fino a $m = 100$, mantenendo h invariato.

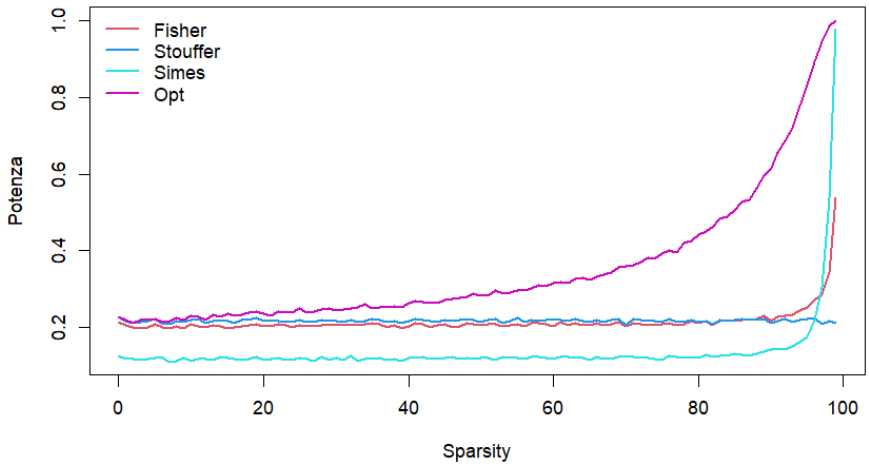

Figura 4.4: Potenza dei metodi con θ_i omogenea, $m = 100 = h = 5$

Il risultato ottenuto appare interessante rispetto all'andamento della potenza dei diversi metodi, piuttosto che individuare quale metodo sia il più potente che ovviamente è il test "ottimo". Infatti, anche in questo caso, i metodi "simil-Bonferroni" sono più potenti quando la *sparsity* è molto elevata, mentre Fisher e Stouffer sono più forti all'aumentare della numerosità dei test non nulli, con un leggero indebolimento del primo rispetto al secondo. Però, è necessario evidenziare come per valori ridotti della *sparsity* l'andamento della potenza per tutti i metodi tende a seguire un andamento costante, ad eccezione di quello ottimo. Tale risultato, probabilmente, è dovuto al fatto che h è rimasto invariato, mentre m è aumentato di molto. Per cui, essendo $\theta_i = h/k$, l'effetto tende a "diluirsi" a causa della numerosità elevata di test non nulli.

Per comprendere meglio quanto detto finora, si prova ad aumentare il segnale $h = 16$, lasciando $m = 100$.

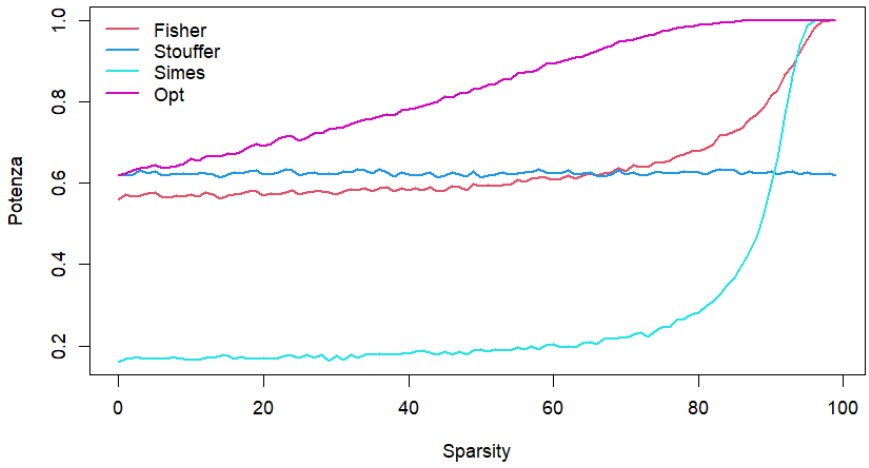

Figura 4.5: Potenza dei metodi con θ_i omogenea, $m = 100 = h = 16$

Quanto ottenuto da quest'ultima simulazione, conferma il fatto che con un'elevata numerosità di test, all'aumentare del numero di test non nulli la potenza dei metodi "non ottimi" tende ad un andamento costante, purché vi sia un adeguato valore del segnale h che permetta di evidenziare tale andamento. Infine, come si è visto fino ad ora, per una *sparsity* elevata i metodi "simil-Bonferroni" sono più potenti, all'aumentare di k Fisher risulta essere più forte fino ad un certo valore di test non nulli, dove Stouffer inizia a rappresentare il metodo migliore per rifiutare le ipotesi nulle false.

Come è stato fatto fino ad ora, anche per lo scenario con distribuzione di h lineare, si verifica, attraverso varie simulazioni con diversi valori di m e h, se i metodi di Bonferroni, Sidak e Simes hanno risultati simili. Giunti alla medesima conclusione si nota come i risultati ottenuti utilizzando un parametro θ_i che cresca in funzione di $i = 1, \ldots, k$ sono sostanzialmente gli stessi ottenuti con il parametro $\theta_i = h/k$ omogeneo lungo il vettore θ.

Per cui, onde evitare inutili ripetizioni di risultati che, nonostante la diversa genesi, sono appena stati illustrati, per le osservazioni riguardanti il modello lineare si preferisce rinviare a quanto è stato detto per il caso di $\theta_i = h/k$.

4.3 Risultato generale

In sintesi, se si volessero riassumere le osservazioni fin qui svolte e fare un confronto tra i diversi scenari, si potrebbe evidenziare che, nonostante sia cambiata la numerosità e il modo di considerare il valore assunto da θ_i, nei casi di *sparsity* massima, ossia $k = 1$, i metodi di Bonferroni, Sidak e Simes sono quelli che garantiscono una potenza maggiore. Mentre per valori di k maggiori Fisher e Stouffer risultano più potenti, con Stouffer superiore a Fisher quando $sparsity \to 0$.

Il confronto appena illustrato non tiene volutamente conto del test "ottimale" che si basa sul lemma *Neyman-Pearson* [Neyman & Pearson (1933)] e che ovviamente rappresenterebbe il metodo più potente in ogni scenario fin qui analizzato. Tuttavia, tale test è stato incluso nella simulazione col fine di evidenziare quale sarebbe in termini teorici la potenza massima raggiungibile da un metodo combinatorio di *p-value* indipendenti.

Inoltre, è stato possibile osservare come all'aumentare di m la potenza dei metodi tende ad avere un andamento sostanzialmente costante quando la *sparsity* diminuisce, ossia quando il numero di test non nulli aumenta.

Nel prossimo capitolo, si discuteranno i risultati ottenuti, cercandone una giustificazione teorica e confrontandoli con quanto la letteratura suggerisce rispetto alla potenza dei metodi combinatori in condizioni di "sparsità".

Capitolo 5

Discussione, limiti e sviluppi futuri

5.1 Discussione

I metodi di combinazione di *p-value* indipendenti utilizzati per testare ipotesi nulle globali sono stati oggetto di studio per molto tempo. In un'opera estremamente rilevante, Birnbaum (1954) ha mostrato che sotto l'assunzione ragionevole e poco restrittiva di distribuzione non crescente dei *p-value* sotto l'alternativa, ogni funzione combinatoria che soddisfa la proprietà della monotonicità (vedi Definizione 5.1) è ottima rispetto ad alcune alternative. Di conseguenza, non c'è un unico test migliore per tutte le alternative in generale.

Definizione 5.1. (**Monotonicità della funzione combinatoria**)
Una funzione combinatoria è detta monotona, se, quando H_0 è rifiutata per una serie di *p-value* $\{p_1, \ldots, p_m\}$, essa sarà ugualmente rifiutata per tutti i *set* $\{p_1^*, \ldots, p_m^*\}$ dove $p_i^* \leq p_i$ per $i = 1, \ldots, m$.

Questo risultato, che escludeva l'ottimalità di un metodo su altri, ha motivato le ricerche di metodi alternativi da usare per definire ed identificare un metodo effettivamente ottimo. Per esempio, si è cercato di rispondere a tale quesito osservando il comportamento dei metodi quando la grandezza campionaria diventa notevole. Un principio di questo tipo è la c.d. *Bahadur efficiency* [(Bahadur, 1967)] che ha prodotto alcuni risultati interessanti. Tuttavia, questo filone di ricerca ha anche evidenziato i suoi limiti quando è stato dimostrato che un certo numero di metodi hanno la stessa efficienza (di Bahadur) e, perciò, sono indistinguibili sotto questo aspetto. La presente tesi si focalizza su un numero finito di statistiche campionarie, per cui questi principi non verranno approfonditi ulteriormente.

Nel contesto di un modello normale, preso a riferimento nello studio di simulazione come illustrato nel Capitolo 3, i metodi di Bonferroni (così come Simes e Sidak) e

Fisher hanno assunto il massimo della loro potenza quando l'effetto è concentrato in un unico test (*sparsity* massima). Questa osservazione empirica è in linea con il risultato teorico derivante dalla funzione di potenza dei metodi. Ossia, si può dimostrare che la funzione della potenza di entrambi i metodi è *Schur convessa* (vedi pp. 79 ss. in Olkin & Marshall (2014)) rispetto al vettore di parametri $\theta = (\theta_1, \ldots, \theta_m)$ e quindi, per $\sum_{i=1}^{m} \theta_i$ fisso, la potenza massima si ottiene con $(\sum_{i=1}^{m} \theta_i, 0, \ldots, 0)$, che corrisponde ad un effetto concentrato in un singolo test. Dall'altro lato, è facilmente dimostrabile che la funzione della potenza del metodo di Stouffer è in funzione solamente di $\sum_{i=1}^{m} \theta_i$, e quindi non dipende dall'allocazione dei segnali. I risultati illustrati nel presente elaborato sono allineati in larga misura con quelli riportati in Loughin (2004), che presenta un confronto sistematico tra i metodi di combinazione. Un'importante differenza riguarda l'osservazione di Loughin per cui la configurazione parametrica più difficile per i metodi di Bonferroni e Fisher è quella con uno o pochi segnali. Ciò è esattamente l'opposto di quanto è stato osservato nei risultati della simulazione svolta. Tale discrepanza può essere giustificata dal differente meccanismo di generazione dei dati; Loughin (2004) considera una distribuzione *Beta* sotto l'alternativa, mentre in questa tesi si è considerata una distribuzione *Normale*. Questo sottolinea la necessità di considerare il tipo di ipotesi alternative nel fornire indicazioni pratiche sulla potenza attesa dai diversi metodi.

Inoltre, nel lavoro di Birnbaum (1954) si esalta il metodo di Fisher, considerandolo il primo (non il migliore) tra gli altri. Tale conclusione appare coerente con quanto ottenuto nel presente studio. Infatti, nonostante certamente non possa classificarsi come il "migliore" per i motivi fin qui illustrati, il metodo di Fisher permette di valorizzare la relazione tra livello di *sparsity* e potenza, limitatamente al modello utilizzato in questa sede.

Nello specifico, il metodo Bonferroni è rinomato per la sua (forse fin troppa [Nakagawa (2004)]) conservatività, anche se, come si è visto, esso garantisce una potenza prossima al massimo in corrispondenza del livello più alto di *sparsity*; potenza che diminuisce pesantemente all'aumentare dei numero di parametri non nulli. Comportamento che si replica anche col metodo di Sidak e di Simes. Per cui, in questi casi, per mantenere la potenza del test sarà necessario propendere per il metodo di Stouffer o quello di Fisher, che, se posti a confronto, appare evidente il vantaggio nell'utilizzo di quest'ultimo. Infatti, dove il metodo di Stouffer garantisce un potenza fissa indipendente dalla *sparsity*, quello di Fisher consente anch'esso di mantenere un livello di potenza minimo (leggermente inferiore a Stouffer) in caso di

bassa concentrazione dell'effetto. Però, allo stesso modo di Bonferroni, esso sfrutta l'aumento della *sparsity*, assumendo la sua potenza massima.

In conclusione, per riprendere la relazione *sparsity*-potenza sopra presentata, il metodo di Fisher risulta essere quello che meglio la sfrutta, comportandosi in maniera simile a Bonferroni quando la concentrazione dell'effetto è alta e come Stouffer se l'effetto è "sparso". Anche se, occorre ricordare, non esiste un metodo migliore degli altri *in toto*, si può concludere che Fisher è il metodo più robusto, limitatamente al modello qui considerato.

5.2 Limiti e sviluppi futuri

Fin dalle prime pagine del presente elaborato si è evidenziato come la simulazione svolta si fondasse su alcune assunzioni e semplificazioni che potrebbero in qualche modo limitare la credibilità dei risultati. Nonostante ciò, prima di compiere un breve *excursus* di tali limiti, si ritiene che quanto ottenuto in questa tesi, oltre a rispondere alla domanda statistica operativa, possa essere un *incipit* per eventuali sviluppi futuri, in cui magari rilassare le assunzioni svolte.

I limiti dello studio riguardano principalmente quattro macro-categorie, ossia l'indipendenza dei *p-value*, il modello scelto, il tipo di alternativa e il modo di definizione dei parametri *input* della simulazione. Nello specifico:

- Assunzione di indipendenza; essa rappresenta l'assunzione più forte tra quelle svolte, in quanto si esclude totalmente che tra i test (ed i rispettivi *p-value*) vi sia una qualsiasi tipologia di dipendenza. A tal proposito, un possibile sviluppo futuro potrebbe essere proprio il fatto di considerare la non indipendenza dei *p-value* [Gasparin et al. (2025)]. In particolare, per il metodo di Fisher, che nella sua "forma" originale richiede tale assunzione, alcuni studiosi hanno proposto delle estensioni che permettano di considerare il caso in cui tale condizione non venga rispettata [Brown (1975), Kost & McDermott (2002)].

- Modello su cui si basa la simulazione; come è già stato evidenziato nella discussione, un limite del modello alla base della simulazione è il fatto che la distribuzione dei dati generati segua una distribuzione *Normale*, il che porta ad una semplificazione, ma anche a possibili distorsioni nei risultati rispetto ad altre distribuzioni [Loughin (2004)]. Inoltre, va evidenziato come la molteplicità dei test venga "ridotta" in un'unica ipotesi nulla globale, la quale comunque

viene spesso usata in tale ambito [Futschik et al. (2019)], nonostante il risultato finale (Rifiuto o meno) non enfatizzi le caratteristiche dei singoli test.

- Tipo di alternativa; ossia il fatto che si è considerato solamente il test unilaterale a coda destra. Tale limite è facilmente ovviabile cambiando i segni per quanto riguarda il caso di alternativa sinistra, ottenendo un risultato sostanzialmente identico. Mentre, occorrerà interrogarsi su cosa accade se si decidesse di utilizzare come riferimento un test bilaterale, in particolare in termini di direzioni dell'effetto [Djordjilović et al. (2026)].

- Definizione dei parametri *input*; i quali sono sì stati selezionati arbitrariamente, ma pur sempre in base ai risultati che si ottenevano nella simulazione con l'inserimento di essi. Nello specifico, si è cercato di individuare le combinazioni di parametri che meglio rispondessero alla domanda statistica operativa, mantenendo fissi quei parametri che non erano direttamente oggetto di quest'ultima $(\alpha\,; B)$. Effettuando tale selezione, inevitabilmente non sono state considerate alcune casistiche che nonostante non siano esplicative in questa situazione, potrebbero esserlo in altre, ossia con domande statistiche differenti. Inoltre, precisamente all'allocazione dell'effetto lungo il vettore θ, occorre ricordare come siano state svolte delle semplificazioni nel caso di $m > 2$. Ossia il fatto di porre il valore di θ_i uguale ad una formula che rispecchiasse un preciso andamento, non considerando anche altre possibili modalità di allocazione di h. Si aggiunge che, come si è evidenziato nel Capitolo 4, entrambe le semplificazioni svolte con $m > 2$ hanno portato a risultati pressoché uguali e, quindi, un possibile spunto per sviluppi futuri potrebbe consistere nell'approfondire le ragioni alla base di quanto ottenuto.

Conclusioni

La simulazione svolta, pur con i limiti illustrati, è in grado di giungere a risultati degni di nota. Infatti, il presente studio, limitatamente al modello considerato, è in grado di comprendere il comportamento dei diversi metodi di combinazione dei *p-value* indipendenti al variare della numerosità m dei test , grandezza dell'effetto e della *sparsity*. Nello specifico, tenendo conto anche del *benchmark* teorico del test "ottimo", si evidenzia come il metodo più robusto nei casi qui analizzati sia quello di Fisher, il quale garantisce una buona potenza al variare dei due elementi sopra citati.

Per cui, tenuto conto che i risultati della simulazione sono limitati dal modello scelto e dalle assunzioni svolte, grazie a quanto ottenuto si è in grado di rispondere alla domanda statistica operativa sul comportamento della potenza dei metodi e, quindi, raggiungere l'obiettivo prefissato all'inizio dello studio.

In conclusione, volendo riassumere quanto è stato appreso nel corso del presente elaborato e focalizzandosi solo sui risultati più rilevanti, si può evidenziare come i metodi di Bonferroni, Sidak e Simes (identificabili come metodi "simil-Bonferroni") siano molto potenti qualora la *sparsity* sia molto grande, ma non si può dire altrettanto quando questa diminuisce, dove invece la loro potenza si riduce di molto. Mentre, per quanto riguarda Stouffer, nella discussione si è già sottolineato come questo metodo non venga influenzato dall'allocazione dell'effetto h e, quindi, sia indifferente al livello di *sparsity*. Infine, passando al metodo di Fisher, esso garantisce una potenza che, pur non essendo sempre la maggiore, mantiene un buon livello, in particolare al variare del livello di *sparsity*, in quanto il suo comportamento è simile ai metodi simil-Bonferroni (se *sparsity* è massima) e al metodo di Stouffer (se k aumenta).

Quanto finora detto non significa che il metodo di Fisher sia migliore degli altri, ma semplicemente esso rappresenta il metodo più robusto che garantisce una buona potenza. Inoltre, le conclusioni appena svolte non sono estendibili in generale, in quanto, occorre ricordare che il modello scelto per la simulazione si basa pur sempre su delle assunzioni (anche molto forti), per cui i risultati ottenuti sono limitati a

specifiche condizioni, qui considerate. Proprio per questo, come si è già fatto intendere nell'elaborato, il punto di partenza per possibili sviluppi futuri è rappresentato da questi limiti che andranno eliminati qualora si volessero generalizzare le evidenze ottenute nel presente studio.

Appendice A

Comandi in R

Pacchetti aggiuntivi da inserire

```
library(grid)
library(gridExtra)
```

A.1 Disegno di simulazione con m=2

```
#Funzione
duetest<-function(m,h,B,alpha){

  theta_grid1<-seq(0, h, by=0.1)
  b<-seq(1,B, by=1)
  p1 <- matrix(NA, nrow = length(theta_grid1), ncol = B)
  p2 <- matrix(NA, nrow = length(theta_grid1), ncol = B)
  test_bon<- matrix(NA, nrow = length(theta_grid1), ncol = B)
  test_fisher<- matrix(NA, nrow = length(theta_grid1), ncol = B)
  test_sidak<- matrix(NA, nrow = length(theta_grid1), ncol = B)
  test_stouffer<- matrix(NA, nrow = length(theta_grid1), ncol = B)
  test_simes<- matrix(NA, nrow = length(theta_grid1), ncol = B)
  test_opt<- matrix(NA, nrow = length(theta_grid1), ncol = B)

  for (i in seq_along(theta_grid1)) {
    theta_1<-theta_grid1[i]
    theta_2<-h-theta_grid1[i]
    for (j in seq_along(b)) {
      X1<-rnorm(1,mean = theta_1)
```

```
      X2<-rnorm(1,mean = theta_2)
      p1[i,j]<-1-pnorm(X1)
      p2[i,j]<-1-pnorm(X2)
      test_bon[i,j]<-min(c(p1[i,j],p2[i,j]))<=alpha/m
      test_fisher[i,j]<- -2*sum(log(c(p1[i,j],p2[i,j])))>qchisq(1-alpha, df
          =2*m)
      test_sidak[i,j]<-min(c(p1[i,j],p2[i,j]))<=1-(1-alpha)^(1/m)
      test_stouffer[i,j]<-1/sqrt(m)*sum(qnorm(1-c(p1[i,j],p2[i,j])))>qnorm
          (1-alpha)
      p_ord<-sort(c(p1[i,j],p2[i,j]))
      test_simes[i,j]<-min(m/(1:m)*p_ord)<=alpha
      test_opt[i,j] <-theta_1*X1 + theta_2*X2 > qnorm(1-alpha, sd=sqrt(theta
          _1^2 + theta_2^2))
    }
  }
  pow_bon = apply(test_bon,1,mean)
  pow_fisher = apply(test_fisher,1,mean)
  pow_sidak = apply(test_sidak,1,mean)
  pow_stouffer = apply(test_stouffer,1,mean)
  pow_simes = apply(test_simes,1,mean)
  pow_opt = apply(test_opt,1,mean)

  plot(theta_grid1,pow_bon,xlab = expression(theta[1]), ylab = "Potenza",
       ylim=c(min(pow_bon,pow_fisher,pow_sidak,pow_simes,pow_stouffer,pow_
           opt),
              max(pow_bon,pow_fisher,pow_sidak,pow_simes,pow_stouffer,pow_
                  opt)), type="n")
  lines(smooth.spline(theta_grid1,pow_bon), col=1, lwd=2)
  lines(smooth.spline(theta_grid1,pow_fisher), col=2, lwd=2)
  lines(smooth.spline(theta_grid1,pow_sidak), col=3, lwd=2)
  lines(smooth.spline(theta_grid1,pow_stouffer), col=4, lwd=2)
  lines(smooth.spline(theta_grid1,pow_simes), col=5, lwd=2)
  lines(smooth.spline(theta_grid1,pow_opt), col=6, lwd=2)
  legend("bottomright", c("Bonf","Fisher","Sidak","Stouffer","Simes","Opt"),
       col=1:6, lwd=2,bty = "n")
```

```
  tabella<-data.frame( Val_theta1 = theta_grid1, Bonf= pow_bon, Sidak= pow_
      sidak, Simes= pow_simes, Fisher= pow_fisher, Stouffer= pow_stouffer,
      Opt= pow_opt)

  tabella <- tabella[tabella$Val_theta1 %in% seq(0, max(theta_grid1)/2, by =
       0.1), ]

  grid.newpage()
  grid.table(round(tabella, 3), rows = NULL)}

  #Simulazioni
  duetest(2,2,10^4,0.1)

  #NON illustrate, ma testate
  duetest(2,4,10^4,0.1)
  duetest(2,6,10^4,0.1)
  duetest(2,2,10^4,0.05)
```

A.2 Disegno di simulazione con m>2 omogenea

```
#Funzione
omogenea<-function(m,h,B,alpha){

  k<-seq(1, m, by= 1 )

  test_bon<-matrix(NA, nrow = length(k), ncol = B)
  test_fisher<-matrix(NA, nrow = length(k), ncol = B)
  test_sidak<-matrix(NA, nrow = length(k), ncol = B)
  test_stouffer<-matrix(NA, nrow = length(k), ncol = B)
  test_simes<-matrix(NA, nrow = length(k), ncol = B)
  test_opt <- matrix(NA, nrow = length(k), ncol = B)

  for (i in seq_along(k)) {
    theta_i<-h/k[i]
    for (j in 1:B) {

      z<-rnorm(m) + c(rep(theta_i,k[i]), rep(0,m-k[i]))
```

```
    pval <- 1-pnorm(z)
    test_bon[i,j]<- min(pval) <=alpha/m
    test_fisher[i,j]<- -2*sum(log(pval)) >qchisq(1-alpha, df=2*m)
    test_sidak[i,j]<-min(pval) <=1-(1-alpha)^(1/m)
    test_stouffer[i,j]<-(1/sqrt(m)*sum(z)) >qnorm(1-alpha)
    test_simes[i,j]<-min( (m*sort(pval))/(1:m) ) <=alpha
    test_opt[i,j] <- (1/sqrt(k[i])*sum(qnorm(1-pval[1:k[i]]))) >qnorm(1-
        alpha)
  }
}

pow_bon = apply(test_bon,1,mean)
pow_fisher = apply(test_fisher,1,mean)
pow_sidak = apply(test_sidak,1,mean)
pow_stouffer = apply(test_stouffer,1,mean)
pow_simes = apply(test_simes,1,mean)
pow_opt = apply(test_opt,1,mean)

plot(m-k,pow_bon,xlab = "Sparsity", ylab = "Potenza",
     ylim=c(min(pow_bon,pow_fisher,pow_sidak,pow_simes,pow_stouffer,pow_
         opt),
             max(pow_bon,pow_fisher,pow_sidak,pow_simes,pow_stouffer,pow_
                 opt)), type="n")
lines(m-k,pow_bon, col=1, lwd=2)
lines(m-k,pow_fisher, col=2, lwd=2)
lines(m-k,pow_sidak, col=3, lwd=2)
lines(m-k,pow_stouffer, col=4, lwd=2)
lines(m-k,pow_simes, col=5, lwd=2)
lines(m-k,pow_opt, col=6, lwd=2)
legend("topleft", c("Bonf","Fisher","Sidak","Stouffer", "Simes","Opt"),
    col=1:6, lwd=2,
       bty = "n")

tabella<-data.frame(Sparsity  = rev(m-k), Bonf= rev(pow_bon), Sidak= rev(
    pow_sidak), Simes= rev(pow_simes), Fisher= rev(pow_fisher), Stouffer=
    rev(pow_stouffer), Opt= rev(pow_opt))
```

```
  if (m > 10) {tabella <- tabella[tabella$Sparsity %in% c(seq(0, max(m-k),
      by = m/10),max(m-k)), ]} else{tabella <- tabella}

  grid.newpage()
  grid.table(round(tabella, 3), rows = NULL)
}

    #Simulazioni
    omogenea(5,2,10^4,0.1)
    omogenea(5,5,10^4,0.1)
    omogenea(100,5,10^4,0.1)
    omogenea(100,16,10^4,0.1)

    #NON illustrate, ma testate
    omogenea(5,1,10^4,0.1)
    omogenea(50,10,10^4,0.1)
    omogenea(5,16,10^4,0.1)
    omogenea(10,10,10^4,0.1)
    omogenea(20,20,10^4,0.1)
```

A.3 Disegno di simulazione con m>2 lineare

```
#Funzione
lineare<-function(m,h,B,alpha){

  k<-seq(1, m, by= 1 )

  test_bon<-matrix(NA, nrow = length(k), ncol = B)
  test_fisher<-matrix(NA, nrow = length(k), ncol = B)
  test_sidak<-matrix(NA, nrow = length(k), ncol = B)
  test_stouffer<-matrix(NA, nrow = length(k), ncol = B)
  test_simes<-matrix(NA, nrow = length(k), ncol = B)
  test_opt <- matrix(NA, nrow = length(k), ncol = B)

  for (i in seq_along(k)) {
    theta_i<-(1:i)*(2*h)/(k[i]*(k[i]+1))
    for (j in 1:B) {
```

```
    z<-rnorm(m) + c(theta_i, rep(0,m-k[i]))
    pval<- 1-pnorm(z)
    test_bon [i,j]<-min(pval) <=alpha/m
    test_fisher[i,j]<- -2*sum(log(pval)) >qchisq(1-alpha, df=2*m)
    test_sidak[i,j]<-min(pval) <= 1-(1-alpha)^(1/m)
    test_stouffer[i,j]<-(1/sqrt(m)*sum(z)) >qnorm(1-alpha)
    test_simes[i,j]<-min( (m*sort(pval))/(1:m) )<=alpha
    test_opt[i,j] <- (1/sqrt(k[i])*sum(qnorm(1-pval[1:k[i]]))) >qnorm(1-
        alpha)
  }
}
  pow_bon = apply(test_bon,1,mean)
  pow_fisher = apply(test_fisher,1,mean)
  pow_sidak = apply(test_sidak,1,mean)
  pow_stouffer = apply(test_stouffer,1,mean)
  pow_simes = apply(test_simes,1,mean)
  pow_opt = apply(test_opt,1,mean)

  plot(m-k,pow_bon,xlab = "Sparsity", ylab = "Potenza",
       ylim=c(min(pow_bon,pow_fisher,pow_sidak,pow_simes,pow_stouffer,pow_
           opt),
              max(pow_bon,pow_fisher,pow_sidak,pow_simes,pow_stouffer,pow_
                  opt)), type="n")
  lines(m-k,pow_bon, col=1, lwd=2)
  lines(m-k,pow_fisher, col=2, lwd=2)
  lines(m-k,pow_sidak, col=3, lwd=2)
  lines(m-k,pow_stouffer, col=4, lwd=2)
  lines(m-k,pow_simes, col=5, lwd=2)
  lines(m-k,pow_opt, col=6, lwd=2)
  legend("topleft", c("Bonf","Fisher","Sidak","Stouffer", "Simes","Opt"),
      col=1:6, lwd=2,
         bty = "n")

  tabella<-data.frame(Sparsity  = rev(m-k), Bonf= rev(pow_bon), Sidak= rev
      (pow_sidak), Simes= rev(pow_simes), Fisher= rev(pow_fisher),
      Stouffer= rev(pow_stouffer),
                      Opt= rev(pow_opt))
```

```
  if (m > 10) {tabella <- tabella[tabella$Sparsity %in% c(seq(0, max(m-k),
       by = m/10),max(m-k)), ]} else{tabella <- tabella}

  grid.newpage()
  grid.table(round(tabella, 3), rows = NULL)
}

#Simulazioni
lineare(5,2,10^4,0.1)
lineare(5,5,10^4,0.1)
lineare(100,5,10^4,0.1)
lineare(100,16,10^4,0.1)

#NON illustrate, ma testate
lineare(5,1,10^4,0.1)
lineare(10,5,10^4,0.1)
lineare(50,10,10^4,0.1)
lineare(20,20,10^4,0.1)
lineare(10,10,10^4,0.1)
lineare(100,50,10^4,0.1)
```

Appendice B

Risultati

B.1 Tabella dei risultati con m=2

Val_theta1	Bonf	Sidak	Simes	Fisher	Stouffer	Opt
0	0.662	0.667	0.668	0.649	0.556	0.762
0.1	0.619	0.623	0.624	0.619	0.543	0.723
0.2	0.6	0.605	0.609	0.614	0.558	0.707
0.3	0.558	0.564	0.567	0.582	0.548	0.665
0.4	0.543	0.547	0.553	0.579	0.554	0.642
0.5	0.512	0.518	0.525	0.57	0.558	0.619
0.6	0.492	0.498	0.507	0.556	0.555	0.596
0.7	0.474	0.48	0.487	0.541	0.548	0.576
0.8	0.46	0.467	0.477	0.536	0.551	0.566
0.9	0.453	0.458	0.468	0.534	0.556	0.56
1	0.442	0.449	0.458	0.522	0.544	0.544

Figura B.1: Potenza dei metodi con $m = 2$ *e* $h = 2$

B.2 Tabelle dei risultati con m>2 omogenea

Sparsity	Bonf	Sidak	Simes	Fisher	Stouffer	Opt
0	0.228	0.236	0.242	0.329	0.348	0.348
1	0.238	0.245	0.252	0.342	0.354	0.393
2	0.256	0.264	0.268	0.347	0.346	0.438
3	0.314	0.323	0.33	0.375	0.348	0.55
4	0.516	0.523	0.528	0.475	0.35	0.76

Figura B.2: Potenza dei metodi con θ_i omogenea, $m = 5$ *e* $h = 2$

Sparsity	Bonf	Sidak	Simes	Fisher	Stouffer	Opt
0	0.552	0.56	0.593	0.793	0.831	0.831
1	0.612	0.621	0.652	0.825	0.825	0.887
2	0.734	0.742	0.766	0.872	0.827	0.946
3	0.897	0.901	0.912	0.928	0.826	0.986
4	0.999	0.999	0.999	0.994	0.826	1

Figura B.3: Potenza dei metodi con θ_i omogenea, $m = 5$ *e* $h = 5$

Sparsity	Bonf	Sidak	Simes	Fisher	Stouffer	Opt
0	0.11	0.115	0.117	0.2	0.218	0.218
10	0.113	0.119	0.12	0.208	0.217	0.226
20	0.113	0.117	0.12	0.208	0.214	0.233
30	0.115	0.121	0.122	0.208	0.221	0.255
40	0.116	0.12	0.123	0.202	0.221	0.267
50	0.111	0.116	0.119	0.206	0.212	0.281
60	0.118	0.123	0.124	0.202	0.217	0.314
70	0.112	0.118	0.119	0.213	0.224	0.358
80	0.117	0.122	0.124	0.217	0.221	0.432
90	0.13	0.137	0.136	0.226	0.218	0.62
99	0.976	0.977	0.977	0.546	0.222	1

Figura B.4: Potenza dei metodi con θ_i omogenea, $m = 100$ *e* $h = 5$

Sparsity	Bonf	Sidak	Simes	Fisher	Stouffer	Opt
0	0.156	0.164	0.166	0.562	0.614	0.614
10	0.157	0.164	0.168	0.574	0.627	0.653
20	0.161	0.168	0.173	0.574	0.622	0.692
30	0.162	0.17	0.175	0.585	0.631	0.737
40	0.165	0.172	0.18	0.59	0.627	0.786
50	0.169	0.178	0.182	0.594	0.621	0.839
60	0.183	0.191	0.2	0.608	0.622	0.892
70	0.2	0.208	0.216	0.63	0.625	0.948
80	0.26	0.271	0.285	0.68	0.626	0.988
90	0.536	0.548	0.582	0.801	0.624	1
99	1	1	1	1	0.627	1

Figura B.5: Potenza dei metodi con θ_i omogenea, $m = 100$ *e* $h = 16$

B.3 Tabelle dei risultati con m>2 lineare

Sparsity	Bonf	Sidak	Simes	Fisher	Stouffer	Opt
0	0.231	0.239	0.246	0.333	0.348	0.348
1	0.253	0.261	0.267	0.345	0.351	0.384
2	0.283	0.291	0.299	0.359	0.35	0.447
3	0.335	0.343	0.348	0.383	0.348	0.541
4	0.521	0.531	0.533	0.481	0.351	0.769

Figura B.6: Potenza dei metodi con θ_i lineare, $m = 5$ *e* $h = 2$

Sparsity	Bonf	Sidak	Simes	Fisher	Stouffer	Opt
0	0.625	0.634	0.661	0.824	0.827	0.827
1	0.711	0.72	0.742	0.86	0.835	0.895
2	0.819	0.826	0.84	0.891	0.825	0.943
3	0.935	0.939	0.943	0.946	0.831	0.987
4	0.998	0.998	0.998	0.993	0.827	1

Figura B.7: Potenza dei metodi con θ_i lineare, $m = 5$ *e* $h = 5$

Sparsity	Bonf	Sidak	Simes	Fisher	Stouffer	Opt
0	0.111	0.117	0.119	0.208	0.221	0.221
10	0.113	0.118	0.12	0.207	0.219	0.228
20	0.111	0.118	0.118	0.205	0.22	0.241
30	0.108	0.113	0.115	0.204	0.213	0.244
40	0.109	0.113	0.115	0.202	0.216	0.258
50	0.115	0.122	0.121	0.208	0.22	0.285
60	0.113	0.119	0.12	0.208	0.218	0.32
70	0.114	0.12	0.12	0.204	0.214	0.353
80	0.125	0.131	0.131	0.211	0.213	0.438
90	0.143	0.15	0.153	0.234	0.214	0.616
99	0.975	0.977	0.976	0.553	0.226	1

Figura B.8: Potenza dei metodi con θ_i lineare, $m = 100$ *e* $h = 5$

Sparsity	Bonf	Sidak	Simes	Fisher	Stouffer	Opt
0	0.16	0.167	0.174	0.588	0.637	0.637
10	0.165	0.173	0.177	0.584	0.63	0.66
20	0.16	0.168	0.174	0.588	0.625	0.692
30	0.173	0.179	0.185	0.595	0.634	0.749
40	0.175	0.182	0.187	0.595	0.613	0.775
50	0.187	0.196	0.204	0.611	0.622	0.845
60	0.207	0.214	0.223	0.633	0.619	0.892
70	0.245	0.256	0.269	0.666	0.629	0.946
80	0.36	0.371	0.391	0.717	0.615	0.989
90	0.804	0.811	0.835	0.868	0.628	1
99	1	1	1	1	0.62	1

Figura B.9: Potenza dei metodi con θ_i lineare, $m = 100$ *e* $h = 16$

Bibliografia

Agresti, A., Franklin, C., Klingenberg, B. & Posner, M. (2017). *Statistics The art and science of learning from data*. Pearson, 4th ed.

Anderson, D. R., Sweeney, D. J., Williams, T. A., Camm, J. D., Cochran, J. J., Fry, M. & Ohlmann, J. (2019). *Essentials of statistics for business and economics*. South Western-Cengage Learning, 9th ed.

Bahadur, R. R. (1967). Rates of convergence of estimates and test statistics. *The Annals of Mathematical Statistics* **38**, 303–324.

Banerjee, A., Chitnis, U., Jadhav, S., Bhawalkar, J. & Chaudhury, S. (2009). Hypothesis testing, type i and type II errors. *Industrial Psychiatry Journal* **18**, 127–131.

Bannerman-Thompson, H., Bhaskara Rao, M. & Chakraborty, R. (2012). Multiple testing of hypotheses in biomedical research. In *Handbook of Statistics*, vol. 28. Elsevier, pp. 201–238.

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E.-J., Berk, R., Bollen, K. A., Brembs, B., Brown, L., Camerer, C. et al. (2018). Redefine statistical significance. *Nature human behaviour* **2**, 6–10.

Birnbaum, A. (1954). Combining independent tests of significance. *Journal of the American Statistical Association* **49**, 559–574.

Bittman, R. M., Romano, J. P., Vallarino, C. & Wolf, M. (2009). Optimal testing of multiple hypotheses with common effect direction. *Biometrika* **96**, 399–410.

Bonferroni, C. (1936). *Teoria statistica delle classi e calcolo delle probabilità*. Pubblicazioni del R. Istituto superiore di scienze economiche e commerciali di Firenze. Seeber.

Boole, G. (1847). *The mathematical analysis of logic*. CreateSpace Independent Publishing Platform.

Brown, M. B. (1975). 400: A method for combining non-independent, one-sided tests of significance. *Biometrics* , 987–992.

Casella, G. & Berger, R. L. (2002). *Statistical inference*. Duxbury, 2nd ed.

Cicchitelli, G., D'Urso, P. & Minozzo, M. (2021). *Statistics: principles and methods*. Pearson Italia.

Clarke, G. M. & Cooke, D. (1992). *A basic course in statistics*. Arnold, 3rd ed.

Deisenroth, M. P., Faisal, A. A. & Ong, C. S. (2020). *Mathematics for machine learning*. Cambridge University Press.

Djordjilović, V., Sofer, T. & Dreyfuss, J. M. (2026). Directional replicability: when can the factor of two be omitted. *Statistics & Probability Letters* , 110662.

Donoho, D. L. (2006). For most large underdetermined systems of linear equations the minimal $l1$-norm solution is also the sparsest solution. *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences* **59**, 797–829.

Dunford, N. & Schwartz, J. T. (1988). *Linear operators, part 1: general theory*. John Wiley & Sons.

Feller, W. (1991). *An introduction to probability theory and its applications, Volume 2*, vol. 2. John Wiley & Sons.

Fisher, R. A. (1934). *Statistical methods for research workers.* Oliver and Boyd, Edinburgh and London.

Futschik, A., Taus, T. & Zehetmayer, S. (2019). An omnibus test for the global null hypothesis. *Statistical methods in medical research* **28**, 2292–2304.

Gale, T., Elsen, E. & Hooker, S. (2019). The state of sparsity in deep neural networks.(2019). *arXiv preprint cs.LG/1902.09574* .

Gasparin, M., Wang, R. & Ramdas, A. (2025). Combining exchangeable p-values. *Proceedings of the National Academy of Sciences* **122**, e2410849122.

Geyer, C. J. (1992). Practical markov chain monte carlo. *Statistical science* , 473–483.

Hogg, R. V., McKean, J. W. & Craig, A. T. (2005). *Introduction to mathematical statistics*. Pearson Prentice Hall, 6th ed.

Kaur, P. & Stoltzfus, J. (2017). Type i, II, and III statistical errors: A brief overview. *International Journal of Academic Medicine* **3**, 268–270.

Keele, L. (2015). The statistics of causal inference: A view from political methodology. *Political Analysis* **23**, 313–335.

Koehler, E., Brown, E. & Haneuse, S. J.-P. (2009). On the assessment of monte carlo error in simulation-based statistical analyses. *The American Statistician* **63**, 155–162.

Kost, J. T. & McDermott, M. P. (2002). Combining dependent p-values. *Statistics & probability letters* **60**, 183–190.

Lee, C.-F., Lee, J. C. & Lee, A. C. (2013). *Statistics for Business and Financial Economics*. Springer New York, 3rd ed.

Lehmann, E. L. & Romano, J. P. (2005). *Testing statistical hypotheses*. Springer.

Liu, Y., Liu, Z. & Lin, X. (2024). Ensemble methods for testing a global null. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **86**, 461–486.

Loughin, T. M. (2004). A systematic comparison of methods for combining p-values from independent tests. *Computational statistics & data analysis* **47**, 467–485.

McCullagh, P. & Polson, N. G. (2018). Statistical sparsity. *Biometrika* **105**, 797–814.

Montgomery, D. C. & Runger, G. C. (2010). *Applied statistics and probability for engineers*. John wiley & sons.

Mukhopadhyay, N. (2000). *Probability and Statistical Inference*. Statistics: a Series of Textbooks and Monographs. Taylor & Francis Group.

Nakagawa, S. (2004). A farewell to bonferroni: the problems of low statistical power and publication bias. *Behavioral ecology* **15**, 1044–1045.

Newbold, P., Carlson, W. L. & Thorne, B. M. (2013). *Statistics for business and economics*. Always learning. Pearson, 8th ed.

Neyman, J. & Pearson, E. S. (1933). Ix. on the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character* **231**, 289–337.

Noether, G. E. (1991). *Introduction to Statistics: The Nonparametric Way*. Springer Texts in Statistics. Springer.

Olive, D. J. (2014). *Statistical theory and inference*. Springer.

Olkin, I. & Marshall, A. W. (2014). *Inequalities: Theory of Majorization and Its Applications*, vol. 143. Academic Press.

Riesz, F. (1910). Untersuchungen über systeme integrierbarer funktionen. *Mathematische Annalen* **69**, 449–497.

Ross, S. M. (2010). *Introductory Statistics*. Elsevier Science & Technology, 3rd ed.

Rossi, R. J. (2018). *Mathematical statistics: an introduction to likelihood based inference*. John Wiley & Sons.

Rothman, K. J. (2010). Curbing type i and type II errors. *European Journal of Epidemiology* **25**, 223–224.

Sarkar, S. (2008). Generalizing simes' test and hochberg's stepup procedure. *The Annals of Statistics* **36**.

Šidák, Z. (1967). Rectangular confidence regions for the means of multivariate normal distributions. *Journal of the American statistical association* **62**, 626–633.

Simes, R. J. (1986). An improved bonferroni procedure for multiple tests of significance. *Biometrika* **73**, 751–754.

Stang, A., Poole, C. & Kuss, O. (2010). The ongoing tyranny of statistical significance testing in biomedical research. *European Journal of Epidemiology* **25**, 225–230.

Stouffer, S. A., Suchman, E. A., DeVinney, L. C., Star, S. A. & Williams, Robin M., J. (1949). *The American Soldier: Adjustment During Army Life*, vol. 1. Princeton, NJ: Princeton University Press.

YOCCOZ, N. G. (1991). Use, overuse, and misuse of significance tests in evolutionary biology and ecology. *Bulletin of the Ecological Society of America* **72**, 106–111.