METODI STATISTICI PER LA BIOINGEGNERIA (B)

PARTE 7: RIASSUNTO SUI TEST STATISTICI, TEST STATISTICI SU CAMPIONI DI GRANDI DIMENSIONI, CORREZIONE PER TEST MULTIPLI

A.A. 2026-2027

Prof. Enrico Longato

OUTLINE

- Riassunto dei test statistici visti
- L'importanza dell'*effect size* quando si analizzano campioni di grandi dimensioni
- Metodi di correzione per test multipli

RIASSUNTO SUI TEST STATISTICI

Abbiamo visto diversi test statistici parametrici e non parametrici per verificare ipotesi statistiche su uno o due campioni.

- Test statistici su un solo campione
- Test statistici su due campioni indipendenti
- Test statistici su due campioni appaiati

TEST STATISTICI SU UN SOLO CAMPIONE

- Voglio verificare un'ipotesi sul valore centrale della distribuzione dei dati.

Test	Assunzioni	Ipotesi nulla	Funzione Matlab
z test	Campione normale, varianza nota	$H_0: \mu = \mu_0$	ztest
t test	Campione normale, varianza incognita	$H_0: \mu = \mu_0$	ttest
Test dei segni	Campione qualsiasi di dati ordinali	$H_0: m = m_0$	signtest

- Voglio verificare un'ipotesi sulla variabilità della distribuzione dei dati.

Test	Assunzioni	Ipotesi nulla	Funzione Matlab
test chi-quadro	Campione normale	$H_0: \sigma^2 = \sigma_0^2$	vartest

TEST STATISTICI SU DUE CAMPIONI INDIPENDENTI (1/2)

- Voglio verificare se i due campioni indipendenti provengono da distribuzioni aventi lo **stesso valore centrale**.

Test	Assunzioni	Ipotesi nulla	Funzione Matlab
z test a due campioni	Campioni normali, varianze note	$H_0: \mu_1 = \mu_2$	Non disponibile
t test a due campioni	Campioni normali, varianze incognite ma uguali	$H_0: \mu_1 = \mu_2$	ttest2
t test di Welch	Campioni normali, varianze incognite	$H_0: \mu_1 = \mu_2$	ttest2 con l'opzione 'Vartype' 'unequal'
test di Wilcoxon Mann-Whitney	Campioni di valori ordinali con distribuzione che differisce solo per uno shift orizzontale	$H_0: m_X = m_Y$	ranksum

TEST STATISTICI SU DUE CAMPIONI INDIPENDENTI (2/2)

- Voglio verificare se due campioni indipendenti hanno la **stessa distribuzione.**

Test	Assunzioni	Ipotesi nulla	Funzione Matlab
t test a due campioni	Campioni normali, varianze incognite ma uguali	$H_0: \mu_1 = \mu_2$	ttest2
test di Wilcoxon Mann-Whitney	Campioni qualsiasi con valori ordinali	$H_0: F_X = F_Y$	ranksum

- Voglio verificare se i due campioni indipendenti hanno la **stessa variabilità.**

Test	Assunzioni	Ipotesi nulla	Funzione Matlab
test F	Campione normale	$H_0: \sigma_X^2 = \sigma_Y^2$	vartest2

TEST STATISTICI SU DUE CAMPIONI APPAIATI

- Voglio verificare se la differenza tra i due campioni appaiati proviene da una distribuzione avente **valore centrale 0.**

Test	Assunzioni	Ipotesi nulla	Funzione Matlab
t test per dati appaiati	Differenze tra i campioni con distribuzione normale	H_0:la differenza tra i campioni ha media 0	ttest passando in ingresso la differenza tra i campioni
test dei ranghi con segno	Differenze tra i campioni con distribuzione simmetrica, valori ordinali	H_0: la differenza tra i campioni ha mediana 0	signrank passando in ingresso la differenza tra i campioni
test dei segni	Differenze tra i campioni con distribuzione qualsiasi, valori ordinali	H_0: la differenza tra i campioni ha mediana 0	signtest passando in ingresso la differenza tra i campioni

OUTLINE

- Riassunto dei test statistici visti
- L'importanza dell'*effect size* quando si analizzano campioni di grandi dimensioni
- Metodi di correzione per test multipli

ESEMPIO: TEST SULLA MEDIA DI DUE POPOLAZIONI NORMALI CON PICCOLA DIFFERENZA TRA LE MEDIE (1/2)

- **Caso con campioni di media dimensione:** generiamo due campioni, **X** e **Y**, di dimensione **n=100** estratti da popolazioni normali di media μ_1**=3.0** e μ_2**=3.1** e deviazione standard $\sigma_1=\sigma_2=2$. Ipotizziamo di non conoscere i parametri delle distribuzioni, ma di sapere che le varianze sono uguali. Applichiamo il t test a due campioni per testare il sistema di ipotesi:
  - $H_0: \mu_1=\mu_2$
  - $H_1: \mu_1\neq\mu_2$
- Che risposta mi dà il test?
- Quanto vale la differenza tra le medie campionarie dei due campioni (*effect size*)?

```
1   clc; clear all; close all
2   rng(0)
3
4   mu1 = 3; mu2 = 3.1; sigma = 2;
5
6   n = 100;
7   X = mu1+sigma*randn(n,1);
8   Y = mu2+sigma*randn(n,1);
9
10  [~,p] = ttest2(X,Y);
11  disp(['p-value = ' num2str(p)])
12
13  mX = mean(X);
14  mY = mean(Y);
15  disp(['Media campionaria di X: ' num2str(mX)])
16  disp(['Media campionaria di Y: ' num2str(mY)])
17  disp(['Effect size: ' num2str(mX-mY)])
18
```

Command Window

New to MATLAB? See resources for Getting Started.

```
p-value = 0.34398
Media campionaria di X: 3.2462
Media campionaria di Y: 2.9546
Effect size: 0.29153
fx >>
```

ESEMPIO: TEST SULLA MEDIA DI DUE POPOLAZIONI NORMALI CON PICCOLA DIFFERENZA TRA LE MEDIE (2/2)

- **Caso con campioni di grande dimensione:** generiamo due campioni, **X** e **Y**, di dimensione **n=100'000** estratti da popolazioni normali di media $\mu_1=3.0$ e $\mu_2=3.1$ e deviazione standard $\sigma_1=\sigma_2=2$. Ipotizziamo di non conoscere i parametri delle distribuzioni, ma di sapere che le varianze sono uguali. Applichiamo il t test a due campioni per testare il sistema di ipotesi:
  - $H_0: \mu_1=\mu_2$
  - $H_1: \mu_1\neq\mu_2$
- Che risposta mi dà il test?
- Quanto vale la differenza tra le medie campionarie dei due campioni (*effect size*)?
- Si tratta di una differenza significativa dal punto di vista pratico?

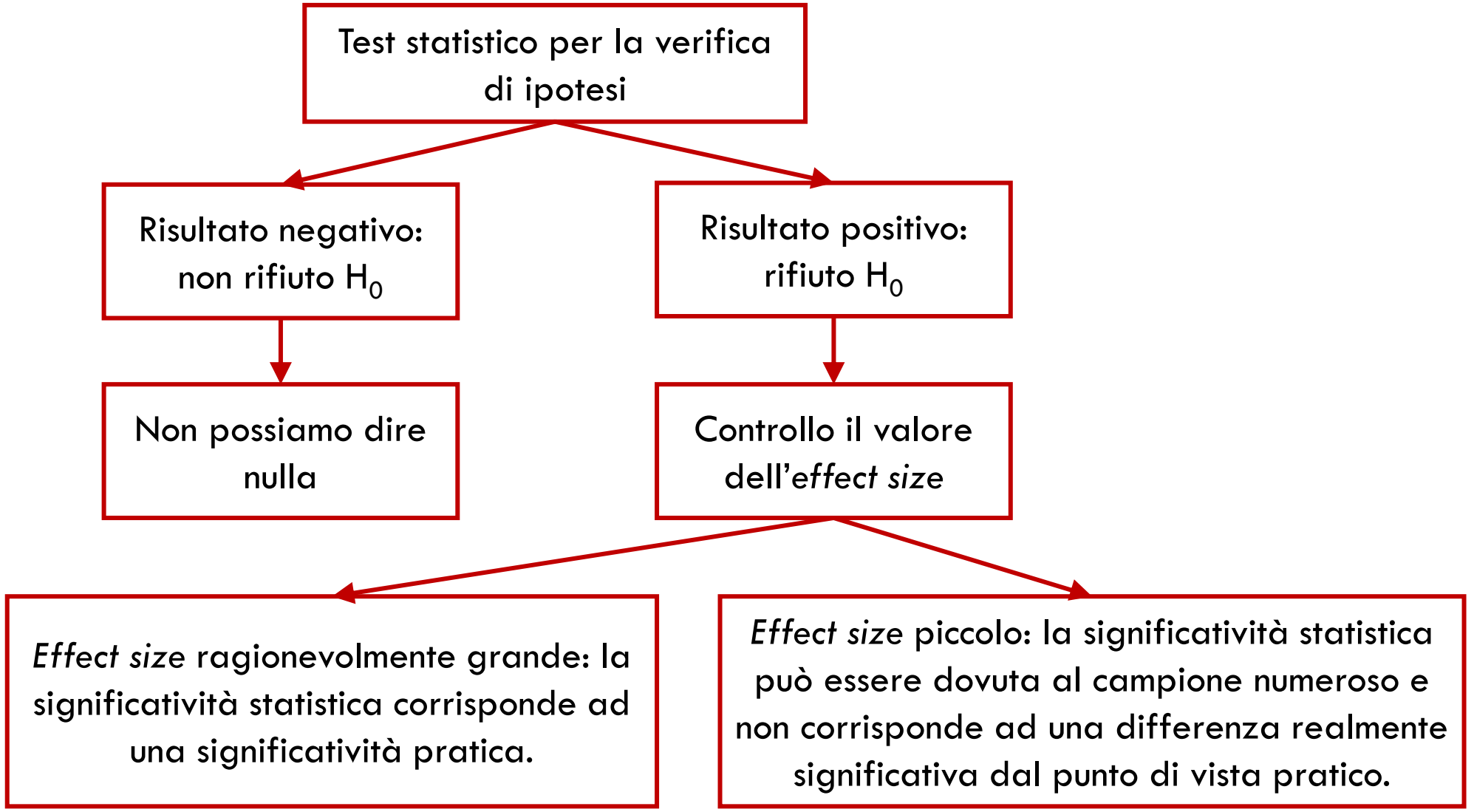
```
1   clc; clear all; close all
2   rng(0)
3
4   mu1 = 3; mu2 = 3.1; sigma = 2;
5
6   n = 100000;
7   X = mu1+sigma*randn(n,1);
8   Y = mu2+sigma*randn(n,1);
9
10  [~,p] = ttest2(X,Y);
11  disp(['p-value = ' num2str(p)])
12
13  mX = mean(X);
14  mY = mean(Y);
15  disp(['Media campionaria di X: ' num2str(mX)])
16  disp(['Media campionaria di Y: ' num2str(mY)])
17  disp(['Effect size: ' num2str(mX-mY)])
18
```

```
Command Window
New to MATLAB? See resources for Getting Started.
  p-value = 7.0678e-30
  Media campionaria di X: 2.9984
  Media campionaria di Y: 3.0999
  Effect size: -0.1015
fx >>
```

TEST STATISTICI CON GRANDI CAMPIONI

- Più i campioni analizzati sono numerosi, maggiore sarà la potenza del test statistico, ovvero la capacità del test di rilevare differenze significative quando effettivamente presenti.
- Se applichiamo un test statistico a campioni molto grandi, il test avrà sufficiente potenza per rilevare come significative anche differenze molto piccole.
- In caso di campioni grandi, occorre quindi prestare attenzione non solo alla significatività statistica ma anche alla significatività pratica della differenza osservata (*effect size*).

VERIFICA DELLA SIGNIFICATIVITA' STATISTICA E VALUTAZIONE DELL'EFFECT SIZE



OUTLINE

- Riassunto dei test statistici visti
- L'importanza dell'*effect size* quando si analizzano campioni di grandi dimensioni
- Metodi di correzione per test multipli

ESEMPIO

Si vuole valutare la capacità di un nuovo algoritmo di pancreas artificiale di ridurre le ipoglicemie nei pazienti diabetici. Si progetta uno studio longitudinale in cui un gruppo di pazienti viene trattato per 6 mesi con la terapia convenzionale (periodo A), e poi per un periodo di altri 6 mesi con il nuovo algoritmo (periodo B). La glicemia dei pazienti viene tracciata mediante un sensore per il monitoraggio in continua della glicemia (CGM). Al termine di ogni fase dello studio si quantificano le ipoglicemie dei pazienti mediante tre indicatori:

- Percentuale di tempo speso in ipoglicemia
- Numero di eventi ipoglicemici
- Durata media degli eventi di ipoglicemia

ANALISI STATISTICA CON CONFRONTI MULTIPLI

End-point dello studio calcolati per ogni paziente
Percentuale di tempo speso in ipoglicemia
Numero di eventi ipoglicemici
Durata media degli eventi di ipoglicemia

Per ciascun end-point si effettua un test statistico per campioni appaiati per confrontare i valori dell'end-point durante il periodo A e B.

Se si osserva una riduzione significativa di almeno uno degli end-point tra il periodo A e B, si conclude che l'algoritmo di pancreas artificiale è efficace nel ridurre le ipoglicemie.

L'esito dell'analisi statistica è basato su **confronti multipli,** ovvero un certo numero $m>1$ di test statistici per la verifica di m ipotesi statistiche (in questo caso: $m=3$).

IL PROBLEMA DEI CONFRONTI MULTIPLI

- Si effettua un'analisi statistica basata su m confronti (m test statistici).
- Ciascuno dei test statistici verifica un'ipotesi statistica con livello di significatività α.
  - La probabilità che il singolo test rifiuti l'ipotesi nulla quando invece essa è vera (errore di prima specie, ossia falso positivo) è pari ad α.
- **Problema**: se l'esito dell'analisi complessiva si basa sull'esito di m test statistici, ciascuno con livello di significatività α, la probabilità che l'analisi complessiva produca un falso positivo è ben maggiore di α !
  - L'analisi complessiva risulta in un falso positivo se almeno uno tra gli m test produce un falso positivo.
  - Tanti più confronti facciamo, tanto è più probabile che almeno uno di essi risulti in un falso positivo.

METODI DI CORREZIONE PER I TEST MULTIPLI

- Metodi che controllano il ***family-wise error rate***
  - Metodo di Bonferroni
  - Metodo di Holm-Bonferroni

- Metodi che controllano il ***false discovery rate***
  - Metodo di Benjamini-Hochberg

- Esistono anche altri metodi di correzione, ma questi sono i più comuni.

FAMILY-WISE ERROR RATE

- ***Family-wise error rate* (FWER), $\bar{\alpha}$:** probabilità che almeno uno degli m test dia un falso positivo, ovvero ci sia almeno un esito positivo quando in realtà tutte le ipotesi nulle testate sono vere → probabilità dell'errore di prima specie dell'analisi complessiva

- Se gli m test sono indipendenti:

$$\bar{\alpha} = 1 - (1 - \alpha)^m$$

Probabilità che tutti gli esiti siano negativi quando le m ipotesi nulle sono tutte vere.

- Se gli m test non sono indipendenti:

$$\bar{\alpha} \leq m \cdot \alpha$$

METODO DI BONFERRONI (1/2)

- Vogliamo che nell'analisi complessiva l'errore di prima specie non superi un certo livello α^*.
- **Idea**: ridurre il livello di significatività utilizzato per i singoli test statistici in modo che il FWER sia al più pari ad α^*.
- Poiché sappiamo che:

$$\bar{\alpha} \leq m \cdot \alpha$$

- Imponiamo:

$$\alpha^* = m \cdot \alpha \quad \rightarrow \quad \alpha = \frac{\alpha^*}{m}$$

- Il valore di α da usare nei singoli test è pari alla probabilità di errore di prima specie che si desidera per l'analisi complessiva diviso il numero di test effettuati.

METODO DI BONFERRONI (2/2)

- Esempio: nell'esempio di slide 14-15, se vogliamo imporre una probabilità di errore di prima specie complessivamente pari al 5%, il metodo di Bonferroni ci suggerisce di usare un livello di significatività nei singoli test pari a $0.05/3 = 0.01\overline{6}$.

- Nota: Il metodo di Bonferroni è conservativo → tende a correggere eccessivamente per cui si ottiene un FWER che in realtà è minore o uguale al valore desiderato → questo a scapito di un aumento dell'errore di seconda specie (falso negativo)

METODO DI HOLM-BONFERRONI (1/2)

- Metodo di correzione che controlla il FWER con minore aumento dell'errore di seconda specie rispetto al metodo di Bonferroni.
- Si ordinano i p-value degli m test dal più piccolo al più grande:

$$p_1 \leq p_2 \leq \cdots \leq p_m$$

- Siano $H_{0,1}, H_{0,2}, \ldots, H_{0,m}$ le m ipotesi nulle corrispondenti a questi p-value.
  - L'ipotesi $H_{0,1}$ viene rifiutata con livello di significatività $\alpha_1 = \frac{\bar{\alpha}}{m}$
  - L'ipotesi $H_{0,2}$ viene rifiutata con livello di significatività $\alpha_2 = \frac{\bar{\alpha}}{m-1}$
  - L'ipotesi $H_{0,3}$ viene rifiutata con livello di significatività $\alpha_3 = \frac{\bar{\alpha}}{m-2}$
  - …

METODO DI HOLM-BONFERRONI (2/2)

- L'i-esima ipotesi nulla viene rifiutata con livello di significatività:

$$\alpha_i = \frac{\bar{\alpha}}{m - i + 1}$$

- Non tutte le ipotesi vengono però effettivamente valutate.
- Si testano le ipotesi a partire da $H_{0,1}$. Appena si trova un'ipotesi $H_{0,i}$ che non viene rifiutata al livello di significatività α_i, si interrompe la procedura.
  - Le ipotesi $H_{0,1}, H_{0,2}, \dots, H_{0,i-1}$ vengono rifiutate.
  - Le ipotesi $H_{0,i}, \dots, H_{0,m}$ vengono accettate.

METODO DI HOLM-BONFERRONI: ESEMPIO

Ipotizziamo che nell'esempio di slide 14-15 si desideri un FWER del 5% e si siano ottenuti i valori di p-value riportati in tabella.

End-point	Ipotesi testata	P-value
Percentuale di tempo speso in ipoglicemia	$H_{0,1}$: Il tempo speso in ipoglicemia non differisce tra periodo A e B.	0.01
Numero di eventi ipoglicemici	$H_{0,2}$: Il numero di eventi ipoglicemici non differisce tra periodo A e B.	0.03
Durata media degli eventi di ipoglicemia	$H_{0,3}$: La durata media degli eventi non differisce tra periodo A e B.	0.06

1) Testiamo $H_{0,1}$ con $\alpha_1 = \frac{0.05}{3} = 0.01\overline{6}$ → esito positivo, rifiutiamo $H_{0,1}$
2) Testiamo $H_{0,2}$ con $\alpha_2 = \frac{0.05}{2} = 0.025$ → esito negativo, accettiamo $H_{0,2}$
3) Accettiamo anche $H_{0,3}$

FALSE DISCOVERY RATE

- ***False discovery rate (FDR):*** frazione di *false discoveries*, ovvero frazione di esiti positivi che sono in realtà dei falsi positivi.
- P: numero di test con esito positivo
- FP: numero di test per cui si ha un falso positivo
- TP: numero di test per cui si ha un vero positivo

$$FDR = \frac{FP}{P} = \frac{FP}{TP + FP}$$

- Esistono dei metodi di correzione che impongono un limite superiore al FDR (invece che al FWER) → metodo di Benjamini-Hochberg

METODO DI BENJAMINI-HOCHBERG (1/2)

- Desideriamo che l'FDR sia al più pari ad α.
- Si ordinano i p-value ottenuti per gli m test dal più piccolo al più grande:

$$p_1 \leq p_2 \leq \cdots \leq p_m$$

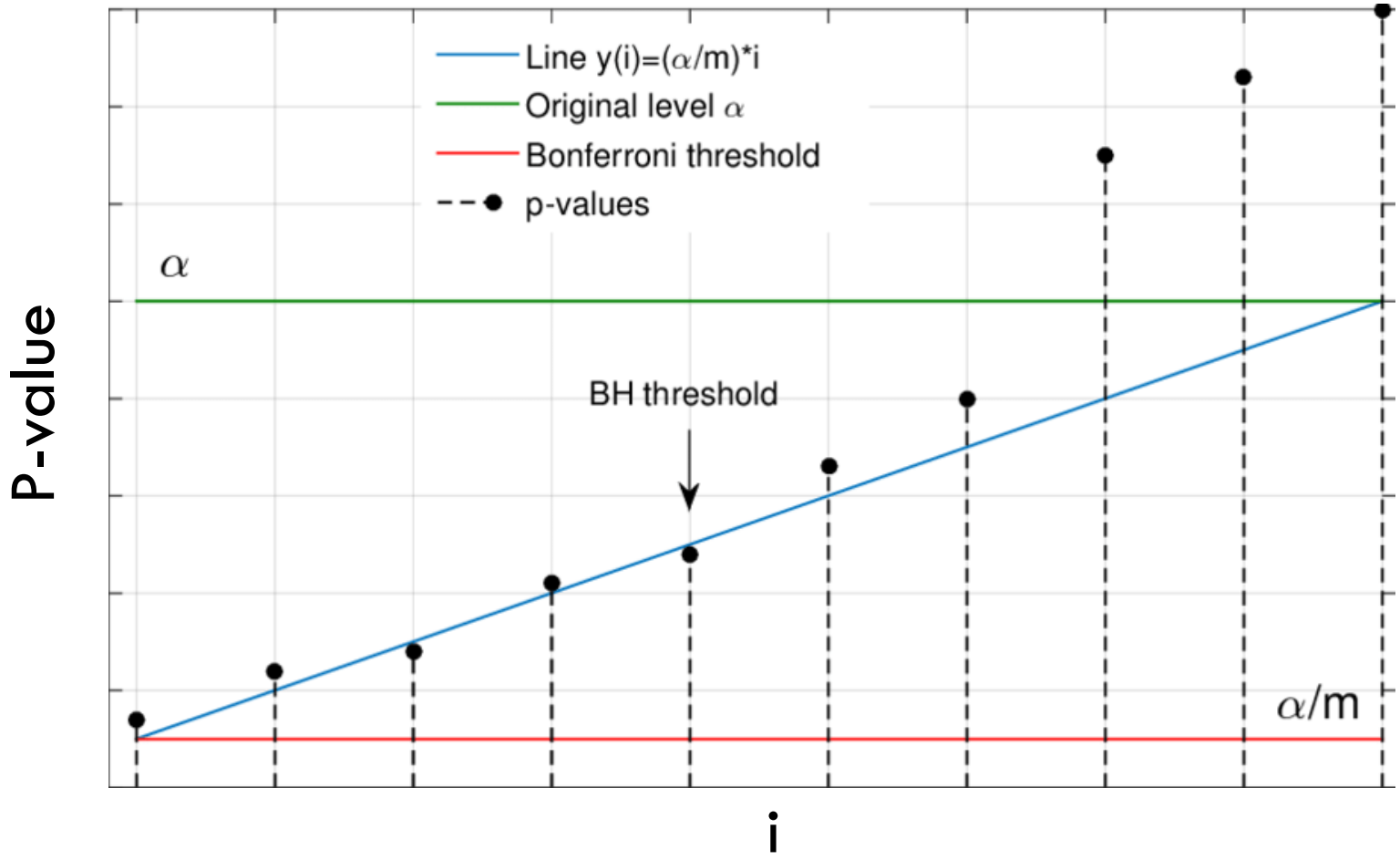
- Siano $H_{0,1}, H_{0,2}, \ldots, H_{0,m}$ le m ipotesi nulle corrispondenti a questi p-value.
- Si cerca il valore di i più grande tale che:

$$p_i \leq i \cdot \frac{\alpha}{m}$$

- Si rifiutano tutte le ipotesi: $H_{0,1}, H_{0,2}, \ldots, H_{0,i}$
- Non si rifiutano le ipotesi: $H_{0,i+1}, \ldots, H_{0,m}$.

METODO DI BENJAMINI-HOCHBERG (2/2)

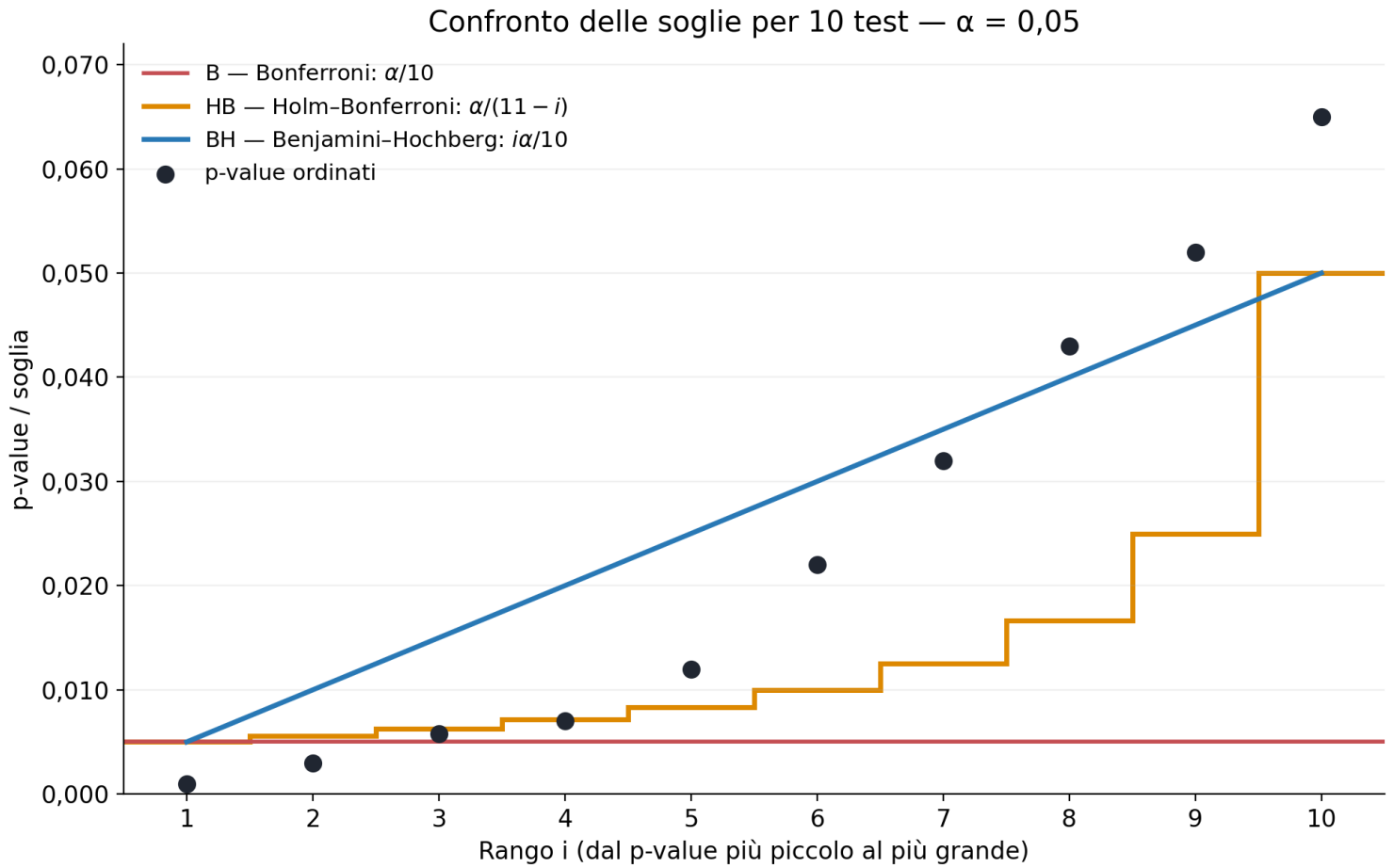
Interpretazione grafica:



CONFRONTO TRA METODI DI CORREZIONE PER TEST MULTIPLI (1 di 2)

Metodo	Quantità controllata	Note
Bonferroni	FWER	Molto conservativo. Errore di prima specie basso, ma aumenta l'errore di seconda specie.
Holm-Bonferroni	FWER	Limita l'errore di prima specie con minore aumento dell'errore di seconda specie rispetto a Bonferroni.
Benjamini-Hochberg	FDR	Tollera maggiore errore di prima specie rispetto ai metodi che controllano FWER. Consigliato quando il numero di confronti da fare è molto alto.

CONFRONTO TRA METODI DI CORREZIONE PER TEST MULTIPLI (2 di 2)



i	p-value	B	HB	BH
1	0,0010	✓	✓	✓
2	0,0030	✓	✓	✓
3	0,0058		✓	✓
4	0,0070		✓	✓
5	0,0120			✓
6	0,0220			✓
7	0,0320			✓
8	0,0430			
9	0,0520			
10	0,0650			

✓ = rifiuto di H_0

QUANDO CORREGGERE PER TEST MULTIPLI

Situazioni tipiche in cui va applicata una correzione per test multipli:

- Studio con end-point multipli
  - Esempio: si vuole valutare l'efficacia di un algoritmo di pancreas artificiale su diversi end-point (tempo in ipoglicemia, numero di eventi di ipoglicemia, ecc.).
- Misure ripetute dell'end-point nel tempo
  - Esempio: si vuole valutare l'impatto di un algoritmo di pancreas artificiale sull'emoglobina glicata e questa viene misurata dopo 3, 6 e 12 mesi di trattamento.
- Studio multibraccio
  - Esempio: si vuole valutare l'efficacia di un algoritmo di pancreas artificiale nel ridurre il tempo speso in ipoglicemia, testando diverse configurazioni dell'algoritmo in parallelo.