La differenza c'è? Si vede?

Elena Spada[1,2], Luigi Gagliardi[1,3], Roberto Buzzetti[1,4]

[1] Laboratorio della Conoscenza Carlo Corchia, Firenze
[2] Biostatistica libero professionista, Milano
[3] Ospedale Versilia, AUSL Toscana Nord-Ovest
[4] Epidemiologo freelance, Bergamo

La storia qui raccontata deriva da uno scenario immaginario, inventato per descrivere brevemente il concetto di significatività statistica e i suoi limiti. In questo contesto si simula una situazione in cui l'efficacia dell'intervento proposto è sconosciuta.
Personaggi:
- il direttore generale: ha il compito di decidere;
- la dottoressa H-0 (H-0): pessimista, secondo lei non c'è nulla che funziona;
- la dottoressa H-1 (H-1): ottimista, vede efficacia in tutti gli interventi;
- l'esperto di statistica: risolve i problemi, o almeno ci prova.

Il direttore generale del distretto di Zamunda ha a cuore il problema di una didattica efficace e inclusiva. Per questo chiede alla sua équipe di proporre un intervento efficace in tal senso con lo scopo di migliorare i risultati scolastici degli studenti. H-1 propone un piano educativo individualizzato (intervento) da svolgere nelle scuole primarie di primo grado.
H-1 propone di sperimentare l'intervento in una scuola primaria scelta casualmente tra tutte le scuole del Distretto. I bambini di questa scuola vengono suddivisi in due gruppi:
1. gruppo di intervento (GI), in cui si introduce l'intervento;
2. gruppo di controllo (GC), in cui si procede con le usuali procedure scolastiche.

L'assegnazione ai gruppi viene fatta facendo sì che le classi dello stesso anno siano distribuite in modo casuale a metà tra GI e GC (randomizzazione stratificata o a cluster).
Un totale di 308 bambini appartengono al GI e un totale di 299 bambini appartengono a GC. Alla fine dell'anno scolastico viene somministrato un test e 4 su 308 bambini del gruppo GI, e 7 su 299 bambini del gruppo GC hanno scarso come risultato [Tabella 1].

Tabella 1. Prima sperimentazione. Distribuzione dei bambini in base ai risultati al test finale

	Risultato al test finale		Rischio di risultato scarso
	soddisfacente	scarso	
GI (intervento)	304	4	4 / 308 = 1,30%
GC (controllo)	292	7	7 / 299 = 2,34%

H-1 calcola il rischio relativo (RR) di ottenere scarso vs soddisfacente: 1,30/2,34 = 0,56 (riduzione del 44%). Dato che RR è inferiore a 1, H-1 conclude che l'intervento è uno strumento efficace nel migliorare i risultati scolastici.
Interviene H-0 sostenendo che in realtà l'intervento non ha alcun effetto e la differenza osservata è solo frutto del caso. A suo parere, se l'intervento fosse inserito in tutte le scuole del distretto non si osserverebbe alcuna efficacia.
Chi tra H-1 e H-0 ha ragione?

Per poter rispondere "senza alcun dubbio", l'unica possibilità sarebbe introdurre l'intervento in tutte le scuole del distretto e valutare, a posteriori, il suo reale effetto. Ovviamente questo non è possibile in una pianificazione sensata dei fondi.
Quindi il direttore generale decide per un approccio conservativo: parte dall'ipotesi che l'intervento non sia efficace (ipotesi H-0), e lascia a H-1 l'onere della prova di efficacia dimostrando "oltre ogni ragionevole dubbio" che H-0 ha torto.
H-1 interpella (tardivamente...) l'esperto di statistica che, utilizzando un test statistico appropriato, calcola la tanto amata, ma spesso fraintesa, "p" ("p-value"). La p quantifica la probabilità di osservare un RR≤0,56 (risultato osservato o più estremo) se l'intervento non è efficace (non c'è differenza tra i gruppi, l'ipotesi H-0 è vera). H-1 potrà sostenere "oltre ogni ragionevole dubbio" che l'ipotesi H-0 non è valida se p risulta "sufficientemente" piccola (rifiuta H-0). Al contrario, non potrà confutare l'ipotesi H-0 (non rifiuta l'ipotesi nulla) per mancanza di prove.
Prima di calcolare p bisogna stabilire la soglia alfa (α) per poter definire p "sufficientemente piccola", cioè per quantificare il valore del "ragionevole dubbio" di cui si parlava. Se $p<\alpha$ la differenza osservata è considerata "statisticamente significativa", permettendo di rifiutare l'ipotesi H-0. α rappresenta il rischio massimo accettabile di dichiarare l'intervento efficace (rifiutando l'Ipotesi H-0) anche se l'intervento, in verità, non è efficace, noto come errore di primo tipo (la differenza non c'è, ma il nostro studio la vede!).
Il direttore stabilisce α=0,05 (5% di rischio di errore di primo tipo) come di consueto. Il test- χ^2 dà come risultato p=0,34, superiore alla soglia 0,05, e si conclude che non è possibile dichiarare l'intervento efficace per mancanza di prove (non si rifiuta l'ipotesi H-0).
H-1 però obietta che p è maggiore di 0,05 perché il numero di soggetti studiati potrebbe essere troppo piccolo, affermando che l'intervento è efficace, ma la potenza (β) dello studio è insufficiente. Si definisce potenza (β) la probabilità di ottenete $p<\alpha$ se, in verità, l'intervento è efficace ed è, quindi, vera H-1. Equivale alla probabilità di dichiarare significativo un test e di rifiutare H-0 se è vera H-1 (la differenza c'è e si vede!).
L'errore di secondo tipo (pari a 1-β) è il rischio di non rifiutare H-0 anche se in realtà è falsa (la differenza c'è, ma il nostro studio non la vede!).
Dunque H-1 chiede, e ottiene la possibilità di ripetere lo studio dopo un adeguato calcolo della numerosità necessaria. Al test di fine anno scolastico 19 su 1569 bambini in GI, e 45 su 1589 bambini in GS hanno risultati scarsi [Tabella 2]: RR=0,43 (riduzione del 57%).

Tabella 2. Seconda sperimentazione. Distribuzione dei bambini in base ai risultati al test finale

	Risultato al test finale		Rischio di risultato scarso
	soddisfacente	scarso	
GI (intervento)	1550	19	19 / 1569 = 1,21%
GC (controllo)	1544	45	45 / 1589 = 2,83%

L'esperto di statistica calcola p=0,0012 (<0,05: se, in verità, l'intervento non è efficace e l'ipotesi H-0 è vera, la probabilità di osservare un RR ≤ 0,43 è 0,0012).
Il direttore, sulla base della significatività statistica, rigetta l'istanza di H-0 (rifiuta l'ipotesi H-0) e conclude che l'intervento è efficace e verrà pertanto introdotto nella didattica di tutte le scuole del distretto.

Riassumiamo

Passi da fare PRIMA di iniziare a raccogliere i dati:
- Stabilire H-0 o ipotesi nulla che è sempre la più conservativa; nel nostro esempio, "l'intervento non è efficace".

- Stabilire H-1 o ipotesi alternativa: "l'intervento è efficace". Durante questa fase occorre quantificare l'entità dell'effetto che aiuta a capire quanto deve essere grande il campione (questione non discussa in questo lavoro).
- Stabilire la potenza desiderata (convenzionalmente 1-β=0,80) e la soglia di significatività, che coincide con il rischio di errore di primo tipo (convenzionalmente α=0,05) per poter calcolare una numerosità campionaria adeguata (questione non discussa in questo lavoro) e poter prendere in seguito una decisione.

Durante l'analisi dei dati:
- usare un test statistico adeguato per poter calcolare p, ossia la probabilità che se è vera l'ipotesi nulla H-0, si osservi un risultato come quello osservato o più estremo.

Quindi:
- se il valore calcolato risulta superiore o uguale alla soglia fissata (p≥α) non si rifiuta l'ipotesi nulla (nel nostro esempio l'intervento non è definito efficace per mancanza di prove);
- se il valore risulta inferiore alla soglia fissata (p<α), si rifiuta l'ipotesi nulla (nel nostro caso l'intervento è definito efficace).

Confronto tra studio e verità. Può succedere che [Tabella 3]:
1. Il trattamento non è efficace e H-0 è vera. In questo caso se:
   – p≥α: H-0 non viene rifiutata, la decisione è aderente con la verità. La probabilità che il campione selezionato per lo studio abbia questo risultato è 1-α;
   – p<α: H-0 viene rifiutata, la decisione non è aderente alla verità e si commette errore di I tipo. La probabilità che il campione selezionato per lo studio abbia questo risultato è α.
2. Il trattamento è efficace e H-0 è falsa. In questo caso se:
   – p≥α: H-0 non viene rifiutata, la decisione è non aderente con la verità e si commette errore di II tipo. La probabilità che il campione selezionato per lo studio abbia questo risultato è β;
   – p<α: H-0 viene rifiutata, la decisione è aderente alla verità. La probabilità che il campione selezionato per lo studio abbia questo risultato è 1-β (potenza).

Tabella 3. Confronto tra realtà e possibili risultati dello studio

		verità	
		H-0 vera (da non rifiutare)	H-0 falsa (da rifiutare)
risultati dello studio	p≥α	corretto	errore di secondo tipo (1-β)
	p<α	errore di primo tipo (α)	corretto

Commento

Cosa succede nella REALTÀ?
Non saremo mai in grado di sapere quale sia l'ipotesi corretta tra "l'intervento è efficace" e "l'intervento non è efficace": la verità resta nascosta. Quello che vediamo (i fenomeni) ci permette di ipotizzare quale sia la verità senza conoscerla, e di prendere decisioni il più possibile fondate. Ciò che conta in medicina non è conoscere la verità (cosa impossibile), bensì prendere delle decisioni razionali, accettando il rischio di sbagliare. L'errore può avere due direzioni:
1. concludere l'inefficacia del trattamento per mancanza di prove anche se nella realtà il trattamento è efficace (errore di II tipo);
2. concludere che il trattamento è efficace anche se nella realtà non lo è (errore di I tipo).

Quando cadiamo in questi errori, trarremo conclusioni "sbagliate", e quanto siamo disposti a rischiare di sbagliare (l'entità del "ragionevole dubbio") lo decidiamo a priori. I valori consueti sono 5% per l'errore di I tipo 20% per l'errore di II tipo. Da notare l'approccio conservativo: si preferisce il rischio di dichiarare come inefficace un trattamento in verità efficace, che quello di dichiarare come efficace un trattamento in verità inefficace.
Quando il risultato dello studio è "significativo", sappiamo quanto è il margine di errore nell'affermare che l'ipotesi nulla è falsa. Al contrario, non significa che l'ipotesi nulla è vera, ma solo che non ci sono prove sufficienti per affermare che l'ipotesi nulla è falsa. Questa sottile differenza è di estrema importanza per poter prendere decisioni. ■

Note

1. **L'esempio utilizzato in questo lavoro è a solo scopo dimostrativo.** I dati sono stati creati *ad hoc* attraverso una simulazione, che permette di avere una popolazione di cui si conoscono i valori "veri" (impossibile nella realtà). La popolazione simulata è costituita da 2.300.000 studenti delle scuole primarie di primo grado (stesso ordine di grandezza delle scuole italiane). La "vera" frequenza di base di bambini con scarsi risultati di apprendimento è stata fissata a 3,19%. che scende a 1,48% dopo intervento (RR=0,46). I due campioni usati per questo esempio sono stati estratti casualmente dalla popolazione simulata. Per avere una potenza β=0,80, una soglia di significatività α=0,05, considerando i valori "veri" dei rischi, e usando il test-χ^2 serve un campione di almeno 1222 bambini per ogni gruppo.
2. **Gli intervalli di confidenza** sono stati volutamente taciuti. Rappresentano un altro metodo per saggiare la significatività e, forse, sarà argomento specifico di un altro articolo.
3. **Attenzione ai test multipli.** Ogni volta che viene eseguito un test si ha un rischio di errore di primo tipo (i.e. di rifiutare l'ipotesi nulla anche se è vera) pari alla soglia fissata. Aumentando il numero di test indipendenti eseguiti in un singolo studio, si aumenta il rischio complessivo di errore di primo tipo (cioè la probabilità di osservare almeno un test significativo anche se è vera l'ipotesi nulla). Per esempio fissando un rischio di errore di primo tipo α=0,05 per ciascun test, se si eseguono due test il rischio che almeno una p sia inferiore a 0,05 è quasi il 10% (9,8%), il 23% se i confronti sono 5 e il 54% per 15 confronti. Per questo, è controindicato eseguire il test quando la risposta è già nota o non è scopo specifico dello studio valutare un certo confronto.
4. **Questo esempio ricostruisce il percorso e gli assunti per gli studi di superiorità** (cioè progettati per tentare di dimostrare la superiorità di un intervento rispetto a uno di controllo). Ultimamente molti studi sono pianificati come studi di uguaglianza o di non-inferiorità; in questi casi il percorso segue modalità diverse.
5. L'approccio che abbiamo descritto in questo lavoro è un **approccio** cosiddetto "classico" o **fisheriano**. In questo metodo si valuta la probabilità di osservare un risultato se l'ipotesi nulla è vera; se risulta molto bassa, si conclude che l'ipotesi nulla non può essere corretta, favorendo l'ipotesi alternativa. Un **approccio** diverso è offerto dalla statistica **bayesiana**, che considera anche la conoscenza a priori. Questo metodo sposta l'inferenza dalla "probabilità dei risultati data l'ipotesi" alla "probabilità dell'ipotesi dati i risultati". Anche questo aspetto meriterà un approfondimento strutturato.

Bibliografia

- Ministero dell'Istruzione. I principali dati relativi agli alunni con DSA. Anno 2022.
- Introduction to p values. In: Harvey Motulsk. Intuitive Biostatistics. Part III. Oxford University Press, 1995.
- Biggeri A. P-value: «Le roi est mort, vive le roi!». Epidemiol Prev. 2019 Mar-Jun;43(2-3):120-1.

elenaspada.bios@gmail.com