Confronto tra due popolazioni

Alessandra Nardi

1 maggio 2021

Indice

Student

Quello che è generalmente noto come t-test, e che più correttamente corrisponde ad una famiglia di test d'ipotesi, nasce nel 1908 quando William Sealy Gosset pubblica, sotto lo pseudonimo di *Student* un articolo dal titolo *The probable error of a mean.* Ve ne raccontiamo brevemente la storia (ma solo per chi conosce l'inglese)

Gosset

Schooled in mathematics and chemistry, Gosset was hired by Arthur Guinness, Son & Co., Ltd. to apply recent innovations in the field of statistics to the business of brewing beer. As a brewer, Gosset analyzed how agricultural and brewing parameters (e.g., the type of barley used) affected crop yields and, in his words, the *behavior of beer*. Because of the cost and time associated with growing crops and brewing beer, Gosset and his fellow *experimental* brewers could not afford to gather the large amounts of data typically gathered by statisticians of their era. Statisticians, however, had not yet developed accurate inferential methods for working with small samples of data, requiring Gosset to develop methods of his own. With the approval of his employer, Gosset spent a year (1906-1907) in the biometric laboratory of Karl Pearson, developing *The probable error of a mean*.

Student

The most immediately striking aspect of *The Probable Error of a Mean* is its pseudonymous author: *Student*. Why would a statistician require anonymity? The answer to this question came publicly in 1930, when fellow statistician Harold Hotelling revealed that *Student* was Gosset, and that his anonymity came at the request of his employer, a *large Dublin Brewery*. At the time, Guinness considered its use of statistics a trade secret and forbade its employees from publishing their work. Only after negotiations with his supervisors was Gosset able to publish his work, agreeing to neither use his real name nor publish proprietary data.

Studio crossover

Esperimento

Sono stati provati due diversi sonniferi (A e B) su un campione di piccole dimensioni di pazienti affetti da insonnia. Ogni paziente ha assunto entrambi i sonniferi, l'ordine di assunzione è stato deciso in modo casuale e per ogni periodo di trattamento è stato registrato il numero medio di ore di sonno guadagnate. Tra i due trattamenti è stato inserito un periodo di *wash out* al fine di eliminare gli effetti del primo farmaco riportando il paziente nelle condizioni iniziali.

Risultati

I risultati sono presentati nella seguenta tabella:

Ore di sonno guadagnate

Pazienti	A	B	B-A
1	+0.7	+1.9	+1.2
2	-1.6	+0.8	+2.4
3	-0.2	+1.1	+1.3
4	-1.2	+0.1	+1.3
5	-0.1	-0.1	+0.0
6	+3.4	+4.4	+1.0
7	+3.7	+5.5	+1.8
8	+0.8	+1.6	+0.8
9	0.0	+4.6	+4.6
10	+2.0	+3.4	+1.4
Media:	$\overline{y}_A = 0.75$	$\overline{y}_B = 2.33$	$\overline{y}_{B-A} = 1.58$

Caratteristiche del disegno sperimentale

- Provando i farmaci sullo stesso paziente tutti i fattori di confondimento legati alle caratteristiche dei soggetti vengono eliminate garantendo una elevata potenza del test che andremo ad eseguire e consentendo di raggiungere spesso risultati conclusivi anche in presenza di campioni di piccola dimensione.
- Il disegno può essere utilizzato solo per patologie croniche per cui è ipotizzabile sospendere il trattamento e ricondurre il paziente in condizioni simili a quelle iniziali.
- Esiste il rischio che nonostante il periodo di wash out l'effetto del primo farmaco si trascini sul secondo. La scelta casuale del primo farmaco consente di verificare la presenza di questo effetto di trascimento che è tuttavia praticamente impossibile correggere.

Obiettivo dello studio

Verificare quale farmaco garantisce la risposta attesa migliore

Assumiamo che la nostra variabile risposta, di natura *continua*, segua (almeno approssimativamente) una distribuzione **normale**.

$$Y_{iA} \sim N(\mu_A, \sigma_A^2)$$

$$Y_{iB} \sim N(\mu_B, \sigma_B^2)$$

Trattandosi di campioni di piccola dimensione questa assunzione è cruciale e deve essere valutata con attenzione.

Nuova variabile D

Trattandosi di risposte osservate sullo stesso paziente è naturale (e necessario) passare a considerare la loro differenza $D_i = Y_{iB} - Y_{iA}$
Dall'ipotesi di normalità su Y_{iA} ed Y_{iB} segue che

$$D_i \sim N(\Delta, \sigma^2_{CO})$$

dove $\sigma^2_{CO} = VAR(Y_{iB} - Y_{iA})$.

Sistema di ipotesi

Il sistema di ipotesi da portare alla verifica dei dati sarà

$$\begin{cases} H_0 : & \Delta = 0 \\ H_1 : & \Delta \neq 0 \end{cases}$$

Statistica test

Costruiamo la nostra statistica test a partire dallo stimatore naturale del valore atteso in un modello normale, cioè la media campionaria

$$\hat{\Delta} = \overline{D} = \frac{\sum D_i}{n}$$

con

$$\overline{D} \sim N(\Delta, \frac{\sigma^2_{CO}}{n}).$$

Standardizzando otteniamo la statistica Z

$$Z = \frac{\overline{D} - \Delta}{\frac{\sigma_{CO}}{\sqrt{n}}}$$

distribuita secondo una normale con valore atteso nullo e varianza unitaria.

Statistica test T

Tuttavia la variabilità della risposta σ^2_{CO} non è generalmente nota ed è necessario stimarla attraverso la *varianza campionaria*

$$S^2_{CO} = \frac{\sum(D_i - \overline{D})^2}{n-1}.$$

Student intuì per primo che la sostituzione di σ^2_{CO} con S^2_{CO} nell'espressione di Z *non è indolore*. Infatti la statistica

$$T = \frac{\overline{D} - \Delta}{\frac{S_{CO}}{\sqrt{n}}}$$

contiene adesso due elementi aleatori, $\overline{D}$ e S_{CO}, che varieranno congiuntamente in ipotetiche ripetizioni dell'esperimento.
Ne segue che T non seguirà più una distribuzione normale.

La t di Student

Student dimostra che, sotto $H_0\,(\Delta = 0)$, abbiamo

$$T \sim t_{(n-1)}$$

dove t indica la famiglia di distribuzioni nota proprio come *t di Student*. Tale famiglia dipende da un solo parametro, i gradi di libertà utilizzati per la stima di σ^2, nel nostro caso $n-1$.
Osserviamo che al crescere di n (e quindi dei gradi di libertà) S^2_{CO} diventerà uno stimatore sempre più preciso di σ^2_{CO} e al tendere di n ad infinito la t di Student convergerà alla densità normale *rimettendo le cose a posto*.
Ma Gosset aveva piccoli campioni ...

Nel nostro esempio:

$$\bar{d} = 1.58 \quad (1)$$

$$s_{CO} = 1.23 \quad (2)$$

$$t_{oss} = \frac{(1.58 - 0)\sqrt{10}}{1.23} = 4.062 \quad (3)$$

Intervallo di confidenza: $(0.70\,,\,2.46)$.

Regione di rifiuto

- Per $\alpha = 0.05$ troviamo nelle tavole $t_{\frac{\alpha}{2},9} = 2.262$
- La regione di rifiuto è $R = \{t : |t| > 2.262\} = (-\infty, -2.262] \cup [2.262, +\infty)$
- Il valore osservato di T appartiene evidentemente alla regione di rifiuto del test conducendo quindi a rifiutare H_0 con una probablità di errore di prima specie pari a 0.05.
  Il p-value è $= 0.0028$: quale informazione aggiuntiva ci fornisce?

Dati indipendenti

Immaginiamo ora che i due sonniferi siano stati assegnati casualmente ai pazienti arruolati nello studio generando due gruppi distinti di soggetti, il gruppo A e il gruppo B.

Il modello

Assumiamo che la variabile risposta ad entrambi i sonniferi segua una distribuzione normale:

$$\begin{array}{rcl} Y_{iA} & \sim & N(\mu_A, \sigma_A),\ i = 1, \dots, n_A \\ Y_{jB} & \sim & N(\mu_B, \sigma_B),\ j = 1, \dots, n_B \end{array}$$

Segue che

$$\begin{array}{rcl} \overline{Y}_A & \sim & N(\mu_A, \frac{\sigma_A}{\sqrt{n_A}}) \\ \overline{Y}_B & \sim & N(\mu_B, \frac{\sigma_B}{\sqrt{n_B}}) \end{array}$$

Le ipotesi

Formalizziamo il problema attraverso il seguente sistema di ipotesi:

$$\begin{cases} H_0: & \mu_A = \mu_B \\ H_1: & \mu_A \neq \mu_B \end{cases}$$

o in modo equivalente:

$$\begin{cases} H_0: & \mu_A - \mu_B = 0 \\ H_1: & \mu_A - \mu_B \neq 0 \end{cases}$$

Sulla variabilità

Nel t-test si assume generalmente che

$$\sigma_A^2 = \sigma_B^2 = \sigma^2.$$

Si tratta di un'ipotesi particolarmente importante dal punto di vista logico poiché implica una identica variabilità nella risposta ai due farmaci che non è affatto scontata e andrebbe sempre preliminarmente verificata.
Facciamo pertanto un test preliminare, il *test F di Fisher*, per verificare se le varianze sono uguali nelle due popolazioni. Le ipotesi a confronto saranno:

$$\begin{cases} H_0: & \sigma_A^2 = \sigma_B^2 \\ H_1: & \sigma_A^2 \neq \sigma_B^2 \end{cases}$$

Consideriamo le varianze campionarie S_A^2 e S_B^2 come stimatori di σ_A^2 e σ_B^2 e definiamo come statistica test il loro rapporto

$$F = \frac{S_A^2}{S_B^2}$$

.

Sotto l'ipotesi nulla di uguali varianze si dimostra che F segue una distribuzione nota come *F di Fisher* che dipende da 2 parametri ν_1 e ν_2. Nel nostro caso tali paramatri uguagliano i gardi di libertà dedicati alla stima delle due varianze, rispettivamente $n_A - 1$ e $n_B - 1$. I simboli

$$F \sim F_{(n_A-1),(n_B-1)}$$

Fissato $\alpha = 0.05$, dovremo costruire la nostra regione di rifiuto che si troverà su entambe le code della distribuzione
Avremo $f_{0.025,9,9} = 0.25$ e $f_{0.975,9,9} = 4.03$ e la nostra regione di rifiuto sarà

$$R = [0, 0.25] \cup [4.03, +\infty).$$

Le nostre stime delle varianze sono $s_A^2 = 3.20$ e $s_B^2 = 4.01$ che conducono a $F_{oss} = 0.80$ che si trova in regione di accettazione (il p-value è 0.7427). Quindi non possiamo rifiutare H_0, si conferma l'assunzione iniziale di uguali varianze.

t test per dati indipendenti

Statistica test

Torniamo al test-t e costruiamo la nostra statistica test a partire da un opportuno stimatore per la quantità di interesse che è adesso $\mu_A - \mu_B$:

Stimatore della differenza dei valori attesi

- $\widehat{\mu_A - \mu_B} = \overline{Y}_A - \overline{Y}_B$

dove $\overline{Y}_A - \overline{Y}_B \sim N(\mu_A - \mu_B, \sigma^2(\frac{1}{n_A} + \frac{1}{n_B}))$.

Stima della varianza

Avendo verificato l'ipotesi di uguali varianze stimiamo la varianza σ^2 come media ponderata di S_A^2 e S_B^2:

$$S^2 = \frac{(n_A - 1)S_A^2 + (n_B - 1)S_B^2}{n_A + n_B - 2}.$$

Sarebbe stato possibile utilizzare la varianza campionaria relativa ad un singolo gruppo?

Statistica test

Definiamo la statistica test

$$T = \frac{\overline{Y}_A - \overline{Y}_B - (\mu_A - \mu_B)}{S\sqrt{\frac{1}{n_A} + \frac{1}{n_B}}}.$$

Sotto l'ipotesi $H_0\,(\mu_B = \mu_B)$

$$T = \frac{\overline{Y}_A - \overline{Y}_B}{S\sqrt{\frac{1}{n_A} + \frac{1}{n_B}}} \sim t_{(n_A+n_B-2)}$$

dove $t_{(n_A+n_B-2)}$ è la distribuzione t con $n_A + n_B - 2$ gradi di libertà.

Esempio

Nel nostro esempio (fittizio):

$$\begin{aligned} \bar{y}_A &= 0.75 \\ \bar{y}_B &= 2.33 \\ s &= 1.898 \end{aligned}$$

Sotto l' ipotesi nulla $\mu_A - \mu_B = 0$:

$$t_{oss} = \frac{2.33 - 0.75}{1.898\sqrt{\frac{1}{10} + \frac{1}{10}}} = 1.861.$$

Fissato $\alpha = 0.05$, avremo $t_{18,0.975} = 2.101$ e la regione di rifiuto:

$$R = \{t : |t| \geq 2.101\} = (-\infty, -2.101] \cup [2.101, +\infty).$$

Adesso non possimo rifiutare H_0 (il p-value è 0.079): cosa è accaduto?

Utilizzando un statistica T per campioni indipendenti quando invece i nostri dati erano appaiati abbiamo sovrastimato la varianza a denominatore (i numeratori delle due statistiche sono identici, perché?) e quindi sottostimato il valore di T.
Infatti

$$Var(\overline{Y}_A - \overline{Y}_B) = Var(\overline{Y}_A) + Var(\overline{Y}_B) - 2Cov(\overline{Y}_A, \overline{Y}_B)$$

L'ultimo addendo, nel caso di dati appaiati, è generalmente positivo riducendo la varianza rispetto al caso di dati indipendenti quando la covarianza è nulla.

Bibliografia

Student (William Sealy Gosset) *The probable error of a mean*. Biometrika 6 (1): 125. March 1908.

Come avremmo proseguito se il test F fosse risultato significativo, rifiutando l'ipotesi di uguali varianze?
Avremmo stimato separatamente le due varianze utilizzando le rispettive varianze campionarie e la statitica test avrebbe assunto la forma

$$T = \frac{\overline{Y}_A - \overline{Y}_B - (\mu_A - \mu_B)}{\sqrt{\frac{S_A^2}{n_A} + \frac{S_B^2}{n_B}}}.$$

La sua distribuzione campionaria sotto l'ipotesi nulla sarebbe stata sempre una t di Student ma per calcolare i gradi di libertà avremmo dovuto usare l'espressione

$$df = \frac{\left(\frac{S_A^2}{n_A} + \frac{S_B^2}{n_B}\right)^2}{\frac{(S_A^2/n_A)^2}{(n_A-1)} + \frac{(S_B^2/n_B)^2}{(n_B-1)}}$$

Il test avrebbe poi seguito gli stessi passi del caso di uguali varianze.

Cosa accadrebbe se non fossimo in grado di ipotizzare un modello normale?
Per grandi campioni saremmo protetti dal teorema centrale del limite che garantisce la convergenza della distribuzione della media campionaria alla densità normale.
Per piccoli campioni dovremmo invece spostarci a considerare metodi non parametrici.

Metodi basati sui ranghi

Alla base di molti metodi non parametrici c'è la nozione di *statistica rango*
Sia $(X_1, \ldots, X_n)$ un campione di osservazioni estratte in modo casuale dalla popolazione d'interesse.
Definiamo *statistica d'ordine* $(X_{(1)}, \ldots, X_{(n)})$ l'insieme delle ossevazioni ordinate in senso crescente, dalla più piccola alla più grande.
La statistica rango $(R_1, \ldots, R_n)$ conterrà invece la posizione che ciascun soggetto occupa nella statistica d'ordine:
$R_i = j \Leftrightarrow X_i = X_{(j)}$
Le due statistiche, tra loro indipendenti, ricostruiscono congiuntamente l'informazione contenuta nel nostro campione.

La distribuzione campionaria della statistica d'ordine dipende dalla distribuzione della variabile oggetto di studio nella popolazione e richiede quindi la definizione di un modello.
Al contrario la distribuzione della statistica rango nel caso di osservazioni i.i.d. è *libera* dalla distribuzione nella popolazione originaria.
Sarà quindi su quest'ultima che baseremo l'analisi quando ci è impossibile assumere un modello, consapevoli tuttavia della inevitabile perdita d'informazione che comporta il fatto di trascurare la statistica d'ordine.
Vediamo come i ranghi verranno utilizzati per costruire la statistica test per il confronto tra due popolazioni nel caso di due campioni indipendenti e nel caso di dati appaiati.

Torniamo allo studio per gruppi paralleli e immaginiamo di aver estratto due campioni casuali dalle due popolazioni a confronto (o di averli generati attraverso l'assegnazione casuale di un trattamento). Per costruire la nostra statistica test calcoliamo statistica d'ordine e rango sul campione congiunto, considerando cioè tutti i dati osservati, indipendentemente dalla popolazione di provenienza o dal trattamento assegnato. Manteniamo tuttavia l'informazione sulla popolazione d'origine e definiamo come statistica test la somma dei ranghi (Wilcoxon rank-sum test) dei soggetti appartenenti ad una delle due popolazioni, in genere quella che corrisponde al campione di dimensione minore che ipottiamo sia il gruppo B.

$$W = \sum_{i \in B} R_i$$

Immaginiamo per semplicità di avere due gruppi di piccola numerosità, $n_A = 5$ e $n_B = 2$ e proviamo a calcolare la distribuzione della statistica W sotto l'ipotesi nulla.
I valori osservabili per i ranghi del campione congiunto andranno da 1 a 6; quali di questi ranghi corrisponderanno alle due osservazioni del gruppo B (R_{1B}, R_{2B})?
I ranghi osservabili saranno

1,2	1,3	1,4	1,5	1,6	1,7
	2,3	2,4	2,5	2,6	2,7
		3,4	3,5	3,6	3,7
			4,5	4,6	4,7
				5,6	5,7
					6,7

A ciascuna coppia di ranghi corrisponderà un valore osservabile di W (la loro somma)

3	4	5	6	7	8
	5	6	7	8	9
		7	8	9	10
			9	10	11
				11	12
					13

Sotto l'ipotesi nulla ognuna di ognuno di questi valori ha la stessa probabilità di essere osservato

La ditribuzione campionaria di W sotto l'ipotesi nulla sarà quindi la seguente

W	Prob
3	1/21
4	1/21
5	2/21
6	2/21
7	3/21
8	3/21
9	3/21
10	2/21
11	2/21
12	1/21
13	1/21

Fissato $\alpha = 0.05$ dovremmo determinare la regione di rifiuto del test su entrambe le code della distribuzione. Notate tuttavia che $1/21 = 0.048 > 0.025$ e di conseguenza la regione di rifiuto del test è vuota. Come mai?
In generale, per piccoli campioni i percentili della distribuzione di W sono ricavabili dalle tavole.

Nel nostro esperimento comparativo sull'efficacia dei due sonniferi, considerando i dati come appartenenti a due campioni indipendenti, otteniamo $W = 130, p - value = 0.06933$, risultato che non consente di rifiutare H_0 assumendo $\alpha = 0.05$.
Notate come l'assenza di un modello, e quindi di parametri che lo identifichino, rende difficile scrivere l'ipotesi nulla e alternativa.
Spesso, per il test della somma dei ranghi di Wilcoxon, si assume come ipotesi nulla l'uguaglianza delle mediane nelle due popolazioni. Tuttavia questa soluzione è valida solo se le due distribuzioni hanno la stessa forma. Più in generale si immagina di estrarre casualmente due soggetti, uno da ogni popolazione. Allora H_0 si configura come

Nel caso di dati appaiati, come per il t-test, passeremo a considerare le differenze D_i. Questa volta calcoleremo le statistiche d'ordine e rango considerando i dati in valore assoluto, ignorando cioè il loro segno. Manterremo tuttavia l'informazione circa il segno delle differenze e definiremo come statistica test la somma dei ranghi corrispondenti a differenze originariamente postive (o negative)

$$S = \sum_{i|D_i>0} R_i$$

Il ragionamneto procede come nel caso precedente osservando che sotto l'ipotesi nulla i ranghi corrispondenti a differenze positive saranno casualmente distribuiti nell'insieme dei ranghi complessivi. Anche in questo caso, fissato α è possibile ricavare dalle tavole i valori dei corrispondenti percentili della distribuzione campionaria di S sotto H_0.

Applicando il test all'esperimento sui due sonniferi (disegno cross-over) otteniamo $S = 55, p - value = 0.009091$, risultato che conduce a rifiutare l'ipotesi nulla con una forte evidenza sperimentale.