Statistica

Esercitazione 15

Alfonso Iodice D'Enza
iodicede@unicas.it

Università degli studi di Cassino

L'importanza del gruppo di controllo

In tutti i casi in cui si voglia studiare l'effetto di un certo fattore, ad esempio l'effetto di una medicina nel trattare una patologia, una condizione desiderabile è che tutti gli altri fattori siano costanti, in modo che ogni variazione dallo stato iniziale (ad esempio lo stato di salute pre-trattamento del paziente) sia ascrivibile unicamente al fattore oggetto di studio.

Poichè una condizione del genere è in molti casi impossibile da ottenere, si procede considerando due campioni:

- un campione viene sottoposto all'effetto del fattore oggetto di studio (ad un gruppo di pazienti viene somministrato il farmaco)
- un altro campione (il gruppo di controllo) non viene sottoposto all'effetto del fattore (al gruppo di controllo dei pazienti, viene somministrato un farmaco del tutto simile a quello in questione, ma senza alcun principio attivo).

Per verificare l'effetto del fattore, si verifica la significatività della differenza di reazione tra i due gruppi.

Si considerino i seguenti due campioni indipendenti

Campione X

Campione casuale $X_1, \ldots, X_n$ da una popolazione normale con media μ_x e varianza σ_x^2.

Campione Y

Campione casuale $Y_1, \ldots, Y_m$ da una popolazione normale con media μ_y e varianza σ_y^2.

Si vuole sottoporre a verifica di ipotesi che le medie delle due popolazioni da cui i campioni sono estratti sono uguali

$$H_0 : \mu_x = \mu_y \quad vs. \quad H_1 : \mu_x \neq \mu_y$$

Gli stimatori utilizzati sono $\bar{X}$ e $\bar{Y}$.

- $\bar{X}$ segue una distribuzione normale con valore atteso $E\left[\bar{X}\right] = \mu_x$ e varianza $var\left(\bar{X}\right) = \frac{\sigma_x^2}{n}$;
- $\bar{Y}$ segue una distribuzione normale con valore atteso $E\left[\bar{Y}\right] = \mu_y$ e varianza $var\left(\bar{Y}\right) = \frac{\sigma_y^2}{m}$.

Verifica Ipotesi su uguaglianza tra medie di popolazioni normali (varianze note)

Per confrontare l'uguaglianza tra le medie si considera lo stimatore

- $\bar{X} - \bar{Y}$ segue una distribuzione normale con valore attesto $E\left[\bar{X} - \bar{Y}\right] = \mu_x - \mu_y$ e varianza $var\left(\bar{X} - \bar{Y}\right) = \frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m}$;
- la versione standardizzata di tale stimatore è dunque

$$\frac{\bar{X} - \bar{Y} - (\mu_x - \mu_y)}{\sqrt{\frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m}}}$$

sotto l'ipotesi nulla $\mu_x = \mu_y$ e dunque $\mu_x - \mu_y = 0$, dunque la statistica test è

$$\frac{\bar{X} - \bar{Y}}{\sqrt{\frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m}}}$$

Verifica Ipotesi su uguaglianza tra medie di popolazioni normali (varianze note): esercizio 1

SI considerino due metodi per realizzare un pneumatico. Il produttore ritiene che non ci sia differenza significativa nella qualità tra gli pneumatici prodotti con i due metodi. Per sottoporre a verifica tale ipotesi vengono realizzate 9 pneumatici con il primo metodo e 7 con il secondo. Il primo campione viene testato su strada su un percorso A, il secondo campione viene testato su un percorso B.
Da studi precedenti è noto che la durata del pneumatico su entrambi i percorsi si distribuisce secondo una normale con la media che dipende dal metodo di costruzione dello pneumatico ma con varianza che dipende dal percorso su cui esso viene testato. In particolare, sul percorso A gli pneumatici hanno una durata caratterizzata da una deviazione standard pari a 3000 km, mentre sul percorso B tale valore è 4000 km.
Sapendo che $\bar{X} = 62.2444$ e $\bar{Y} = 58.2714$, i due metodi producono pneumatici con la stessa durata media? Effettuare una verifica di ipotesi ad una significatività del 5%.

Svolgimento

I dati del problema (in migliaia di km) sono dunque $\bar{X} = 62.2444$, $\bar{Y} = 58.2714$, $n = 9$, $m = 7$, $\sigma_x = 3$ e $\sigma_y = 4$.

$$Z_{oss} = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m}}} = \frac{62.2444 - 58.2714}{\sqrt{\frac{3^2}{9} + \frac{4^2}{7}}} = 2.192$$

il valore critico, ad un livello di significatività $\alpha = 0.05$ è $Z_c = \pm 1.96$: poichè 2.192 è esterno all'intervallo $[-1.96, 1.96]$, si rifiuta H_0.
Il $p.value$ è $2P(Z > z_{oss}) = 0.0284$ (si moltiplica per due perchè l'Hp alternativa è bidirezionale).

Si considerino i seguenti due campioni indipendenti

Campione X

Campione casuale $X_1, \ldots, X_n$ da una popolazione normale con media μ_x e varianza σ_x^2, entrambe incognite. Si assume che la taglia n del campione sia elevata $(n > 30)$, dunque è ragionevole stimare la varianza della popolazione σ_x^2 con lo stimatore varianza campionaria S_x^2.

Campione Y

Campione casuale $Y_1, \ldots, Y_m$ da una popolazione normale con media μ_y e varianza σ_y^2, entrambe incognite. Si assume che la taglia m del campione sia elevata $(m > 30)$, dunque è ragionevole stimare la varianza della popolazione σ_y^2 con lo stimatore varianza campionaria S_y^2.

Si vuole sottoporre a verifica di ipotesi che le medie delle due popolazioni da cui i campioni sono estratti sono uguali

$$H_0 : \mu_x = \mu_y \quad vs. \quad H_1 : \mu_x \neq \mu_y$$

Gli stimatori utilizzati sono $\bar{X}$ e $\bar{Y}$.

- $\bar{X}$ segue una distribuzione normale con valore atteso $E\left[\bar{X}\right] = \mu_x$ e varianza $var\left(\bar{X}\right) = \frac{S_x^2}{n}$;
- $\bar{Y}$ segue una distribuzione normale con

Per confrontare l'uguaglianza tra le medie si considera lo stimatore

- $\bar{X} - \bar{Y}$ segue una distribuzione normale con valore attesto $E\left[\bar{X} - \bar{Y}\right] = \mu_x - \mu_y$ e varianza $var\left(\bar{X} - \bar{Y}\right) = \frac{S_x^2}{n} + \frac{S_y^2}{m}$;
- la versione standardizzata di tale stimatore è dunque

$$\frac{\bar{X} - \bar{Y} - (\mu_x - \mu_y)}{\sqrt{\frac{S_x^2}{n} + \frac{S_y^2}{m}}}$$

sotto l'ipotesi nulla $\mu_x = \mu_y$ e dunque $\mu_x - \mu_y = 0$, dunque la statistica test è

$$\frac{\bar{X} - \bar{Y}}{\sqrt{\frac{S_x^2}{n} + \frac{S_y^2}{m}}}$$

che si distribuisce secondo una normale standard.

Verifica Ipotesi su uguaglianza tra medie di popolazioni normali (varianze non note e campioni di grandi dimensioni): esercizio 2

Si vuole studiare l'efficacia di un nuovo farmaco per ridurre il colesterolo. Per testare il farmaco vengono impiegati 100 volontari, suddivisi in due gruppi da 50. Al primo gruppo viene somministrato il nuovo farmaco; al secondo gruppo, il gruppo di controllo, viene somministrata della lovastatina, sostanza di uso comune per ridurre il colesterolo.
A ciascun volontario è stato detto di prendere una pillola ogni 12 ore per tre mesi. Nessuno dei pazienti sapeva se stesse prendendo il nuovo farmaco o la lovastatina. Il primo gruppo (che ha preso il nuovo farmaco) ha fatto registrare una diminuzione media di colesterolo pari a 8.8, con una variabilità nel campione pari a 4.5.
Il secondo gruppo ha fatto registrare una diminuzione media di colesterolo pari a 8.2, con una variabilità nel campione pari a 5.4.
Questi risultati supportano l'ipotesi che, ad un livello di significatività del 5%, il nuovo farmaco produce in media un decremento maggiore del livello di colesterolo?

Svolgimento

Si vuole sottoporre a verifica $H_0 : \mu_x <= \mu_y$ vs $H_1 : \mu_x > \mu_y$.
I dati del problema sono dunque $\bar{X} = 8.8$, $\bar{Y} = 8.2$, $n = 50$, $m = 50$, $S_x^2 = 4.5$ e $S_y^2 = 5.4$.

$$Z_{oss} = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{S_x^2}{n} + \frac{S_y^2}{m}}} = \frac{8.8 - 8.2}{\sqrt{\frac{4.5}{50} + \frac{5.4}{50}}} = 1.3484$$

il valore critico, ad un livello di significatività $\alpha = 0.05$ è $Z_c = 1.645$: dunque risulta che $Z_{oss} < Z_c$, pertanto non si rifiuta H_0.
Il $p.value$ è $P(Z > z_{oss}) = 0.089$.

Verifica Ipotesi su uguaglianza tra medie di popolazioni normali (varianze non note ma supposte uguali e campioni di piccole dimensioni)

Si considerino i seguenti due campioni indipendenti

Campione X

Campione casuale $X_1, \ldots, X_n$ da una popolazione normale con media μ_x e varianza σ_x^2, entrambe incognite.

Campione Y

Campione casuale $Y_1, \ldots, Y_m$ da una popolazione normale con media μ_y e varianza σ_y^2, entrambe incognite.

In molte applicazioni in cui si è interessati a confrontare le medie di due popolazioni, è ragionevole assumere che σ_x^2 e σ_y^2, sebbene incognite, siano uguali; dunque in questo caso $\sigma^2 = \sigma_x^2 = \sigma_y^2$. Il problema è che si σ^2 non è nota.

stimare σ^2

Si sfrutta il fatto che sia S_x^2 che S_y^2 sono stimatori di σ^2, e quindi si utilizzano entrambi, combinandoli. Si effettua una media ponderata di S_x^2 e S_y^2. Il peso di ciascun stimatore dipende dalla numerosità di ciascun campione, e dai gradi di libertà in particolare. Poichè S_x^2 ha $n-1$ g.d.l. e S_y^2 ha $m-1$ g.d.l., il peso da attribuire a ciascuno stimatore sarà dato dal rapporto tra i g.d.l. dello stimatore e la somma dei g.d.l. complessivi.

$$S_p^2 = S_x^2 \underbrace{\left(\frac{n-1}{n-1+m-1}\right)}_{peso\ x} + S_y^2 \underbrace{\left(\frac{m-1}{n-1+m-1}\right)}_{peso\ y}$$

Per confrontare l'uguaglianza tra le medie si considera lo stimatore

- $\bar{X} - \bar{Y}$ segue una distribuzione normale con valore attesto $E\left[\bar{X} - \bar{Y}\right] = \mu_x - \mu_y$ e varianza $var\left(\bar{X} - \bar{Y}\right) = S_p^2\left(\frac{1}{n} + \frac{1}{m}\right) = \left[S_x^2\left(\frac{n-1}{n-1+m-1}\right) + S_y^2\left(\frac{m-1}{n-1+m-1}\right)\right]\left(\frac{1}{n} + \frac{1}{m}\right)$;
- la versione standardizzata di tale stimatore è dunque

$$\frac{\bar{X} - \bar{Y} - (\mu_x - \mu_y)}{\sqrt{S_p^2\left(\frac{1}{n} + \frac{1}{m}\right)}}$$

sotto l'ipotesi nulla $\mu_x = \mu_y$ e dunque $\mu_x - \mu_y = 0$, dunque la statistica test è

$$\frac{\bar{X} - \bar{Y}}{\sqrt{S_p^2\left(\frac{1}{n} + \frac{1}{m}\right)}}$$

che si distribuisce secondo una t-student con $n - 1 + m - 1 = n + m - 2$ g.d.l.

Verifica Ipotesi su uguaglianza tra medie di popolazioni normali (varianze non note ma supposte uguali e campioni di piccole dimensioni): esercizio 3

Un campione di 22 volontari è stato utilizzato per studiare l'effetto della vitamina C sugli stati influenzali. A dieci volontari è stato somministrato 1 grammo di vitamina C. Ai restanti 12 volontari è stato dato una pillola simile alla vitamina C ma senza principio attivo.
La somministrazione è continuata fino alla sparizione dei sintomi influenzali. Di volta in volta è stato registrato il tempo trascorso dalla prima somministrazione alla guarigione.
Si vuole verificare se la vitamina C contribuisce a ridurre il tempo di guarigione dall'influenza.

gruppo di controllo: $\{6.5, 6, 8.5, 7, 6.5, 8, 7.5, 6.5, 7.5, 6, 8.5, 7\}$
gruppo di studio: $\{5.5, 6, 7, 6, 7.5, 6, 7.5, 5.5, 7, 6.5\}$

Svolgimento

Si vuole sottoporre a verifica $H_0 : \mu_x >= \mu_y$ vs $H_1 : \mu_x < \mu_y$.
I dati campionari portano alle seguenti quantità $\bar{X} = 7.125$, $\bar{Y} = 6.45$, $n = 10$, $m = 12$, $S^2_x = 0.778$ e $S^2_y = 0.581$.

$$S^2_p = S^2_x \left(\frac{n-1}{n-1+m-1} \right) + S^2_y \left(\frac{m-1}{n-1+m-1} \right) = 0.778 \frac{11}{20} + 0.581 \frac{9}{20} = 0.689$$

$$T_{oss} = \frac{\bar{X} - \bar{Y}}{\sqrt{S^2_p \left(\frac{1}{n} + \frac{1}{m} \right)}} = \frac{7.125 - 6.45}{\sqrt{0.689 \left(\frac{1}{12} + \frac{1}{10} \right)}} = 1.9$$

il valore critico, ad un livello di significatività $\alpha = 0.05$ e 20 g.d.l. è $T_c = 1.725$: dunque risulta che $T_{oss} > T_c$, pertanto si rifiuta H_0.
Il $p.value$ è $P(T > t_{oss}) = 0.0375$.

Verifica Ipotesi su uguaglianza tra medie: campioni appaiati

Paired samples

Siano dati due campioni $X_1, \ldots, X_n$ e $Y_1, \ldots, Y_n$ provenienti da popolazioni normali di media μ_x e μ_y rispettivamente. Talora esiste una relazione tra le singole osservazioni X_i e Y_i $(i = 1, \ldots, n)$, dunque i campioni in questione *non* sono indipendenti.
La mancanza di indipendenza trai campioni bisogna procedere diversamente da quanto fatto in precedenza.

Ex.5: Verifica Ipotesi sulla indipendenza tra mutabili

Per testare l'efficacia di un farmaco nella cura di una malattia si considera un campione di 200 pazienti, li si suddivide in due gruppi (A e B). Al gruppo A viene somministrato il farmaco, al gruppo B no; si osserva poi, per ciascun gruppo, il numero di pazienti guariti. I risultati della prova sono riassunti nella seguente tabella

	guariti	non guariti	tot.
gruppo A	75	25	100
gruppo B	65	35	100
tot	140	60	200

Table: Appartenenza ai gruppi vs. guariti (non guariti).

A partire da tale esperimento, si costruisca un test per valutare l'efficacia del farmaco nella cura della malattia.

Svolgimento

Per stabillire se il farmaco sia o meno curativo si può sottoporre la verifica di ipotesi che le due variabili *gruppo di appartenenza* e *guarigione dalla malattia* siano o meno indipendenti: se così dovesse risultare, ovvero che la guarigione dalla malattia non è influenzata dalla assunzione del farmaco, si potrebbe concludere che il farmaco è inefficace.

Ex.5: Verifica Ipotesi sulla indipendenza tra mutabili

Svolgimento

Formalmente

- H_0 : le variabili considerate sono indipendenti.
- H_1 : le variabili considerate non sono indipendenti.

Al fine di calcolare l'indice quadratico di connessione è necessario calcolare le frequenze che ci si attenderebbe se, fissati i marginali di tabella, le variabili fossero indipendenti.

	guariti	non guariti	tot.
gruppo A	70	30	100
gruppo B	70	30	100
tot	140	60	200

Table: Frequenze attese sotto l'ipotesi di indipendenza

Svolgimento

Si passa a calcolare l'indice quadratico di connessione

$$\chi^2 = \sum_{i=1}^{h}\sum_{j=1}^{k}\frac{(n_{ij}-\hat{n_{ij}})^2}{\hat{n_{ij}}} = \frac{(75-70)^2}{70}+\frac{(25-30)^2}{30}+\frac{(65-70)^2}{70}+$$
$$+\frac{(35-30)^2}{30} = 2.38$$

La statistica appena calcolata si distribuisce secondo una v.c. *chi-quadro* χ^2 con $(h-1)\times(k-1)=1$ grado di libertà (h e k sono rispettivamente il numero di modalità delle mutabili considerate). Ad un livello di significatività del 5% e con 1 g.d.l. il valore critico è $\chi^2_{0.95,1}=3.84$. Poichè il valore osservato della statistica è minore del valore critico $(2.38<3.84)$, non si può rifiutare l'ipotesi nulla: di conseguenza non si può sostenere che il farmaco abbia effetti sulla cura della malattia.