ESERCITAZIONE 14/01/19

PROBLEMA A

DATI CATEGORIALI: CONFRONTO TRA 2 PROPORZIONI

Si possono confrontare le probabilità usando le loro differenze $\pi_2 - \pi_1$

→ La differenza tra le proporzioni campionarie stima $\pi_2 - \pi_1$

Se n_1 ed n_2 sono abbastanza grandi, lo stimatore ha una distribuzione campionaria approssimativamente normale.

Qual'è l'errore standard della differenza tra proporzioni campionarie?

$$se = \sqrt{(se_1)^2 + (se_2)^2} = \sqrt{[\hat{\pi}_1(1 - \hat{\pi}_1)]/n_1 + [\hat{\pi}_2(1 - \hat{\pi}_2)]/n_2}$$

INTERVALLO DI CONFIDENZA PER $\pi_2 - \pi_1$

Per grandi campioni casuali indipendenti, un intervallo di confidenza per la differenza $\pi_2 - \pi_1$ tra due proporzioni a livello di popolazione è

stima puntuale ± margine d'errore

$$(\hat{\pi}_2 - \hat{\pi}_1) \pm z\,(se)$$

livello di confidenza 95%, z=1.96

INTERVALLO DI CONFIDENZA PER $\pi_2 - \pi_1$

INTERPRETAZIONE

1. quando l'intervallo di confidenza per $\pi_2 - \pi_1$ **include lo 0**: è plausibile ritenere che $\pi_2 - \pi_1 = 0$, cioè è credibile che $\pi_1 = \pi_2$

   → non vi sono sufficienti evidenze per concludere quale probabilità tra π_1 e π_2 sia più grande
2. quando l'intervallo di confidenza per $\pi_2 - \pi_1$ contiene **solo valori negativi**: π_2 è più piccolo di π_1
3. quando l'intervallo di confidenza per $\pi_2 - \pi_1$ contiene **solo valori positivi**: π_2 è più grande di π_1

TEST DI SIGNIFICATIVITÀ PER $\pi_2 - \pi_1$

1. assunzioni
2. ipotesi
3. test
4. P-valore
5. conclusioni

TEST DI SIGNIFICATIVITÀ PER $\pi_2 - \pi_1$

1. assunzioni

a) casualità → dati ottenuti utilizzando la randomizzazione

b) ≥ 10 osservazioni in ciascuna categoria, per ogni gruppo

TEST DI SIGNIFICATIVITÀ PER $\pi_2 - \pi_1$

2. ipotesi

H_o: $\pi_1 = \pi_2$ $(\pi_2 - \pi_1 = 0)$

H_a: $\pi_1 \neq \pi_2$ $(\pi_2 - \pi_1 \neq 0)$ bidirezionale

oppure H_a: $\pi_1 > \pi_2$ $(\pi_2 - \pi_1 < 0)$ unidirezionale

H_a: $\pi_1 < \pi_2$ $(\pi_2 - \pi_1 > 0)$ unidirezionale

TEST DI SIGNIFICATIVITÀ PER $\pi_2 - \pi_1$

3. test

$z = (\hat{\pi}_2 - \hat{\pi}_1)/se_o$

dove l'errore standard sotto l'ipotesi nulla

$se_o = \sqrt{\hat{\pi}(1 - \hat{\pi})(1/n_1 + 1/n_2)}$

con $\hat{\pi}$ = stima conglobata
si ottiene raggruppando le osservazioni dei 2 campioni

TEST DI SIGNIFICATIVITÀ PER $\pi_2 - \pi_1$

4. P-valore

il P-valore è il peso dell'evidenza contro H_o

5. conclusioni

interpretazione del P-valore → piccoli valori di P forniscono forti evidenze contro H_o e a favore di H_a

Si rifiuta H_o se $P \leq \alpha$ (livello di significatività)

NON si rifiuta H_o se $P \geq \alpha$ (livello di significatività)

PROBLEMA B

DATI QUANTITATIVI: CONFRONTO TRA 2 MEDIE

Si deve distinguere tra:

- campioni indipendenti
- campioni dipendenti (dati appaiati)

PROBLEMA B

CONFRONTO TRA 2 MEDIE - campioni indipendenti

Si possono confrontare le medie usando le loro differenze $\mu_2 - \mu_1$

→ La differenza tra le medie campionarie stima $\mu_2 - \mu_1$

Qual'è l'errore standard della differenza tra medie campionarie?

$$se = \sqrt{(s_1^2/n_1 + s_2^2/n_2}$$

INTERVALLO DI CONFIDENZA PER $\mu_2 - \mu_1$

Per due campioni casuali indipendenti selezionati da una distribuzione normale, un intervallo di confidenza per la differenza $\mu_2 - \mu_1$ è

stima puntuale ± margine d'errore

$$(\bar{y}_2 - \bar{y}_1) \pm t\,(se)$$

INTERVALLO DI CONFIDENZA PER $\mu_2 - \mu_1$

INTERPRETAZIONE

1. quando l'intervallo di confidenza per $\mu_2 - \mu_1$ **include lo 0**: è plausibile ritenere che $\mu_2 - \mu_1 = 0$, cioè è credibile che $\mu_1 = \mu_2$
2. quando l'intervallo di confidenza per $\mu_2 - \mu_1$ contiene **solo valori negativi**: μ_2 è più piccolo di μ_1
3. quando l'intervallo di confidenza per $\mu_2 - \mu_1$ contiene **solo valori positivi**: μ_2 è più grande di μ_1

TEST DI SIGNIFICATIVITÀ PER $\mu_2 - \mu_1$ (campioni indipendenti)

1. assunzioni
2. ipotesi
3. test
4. P-valore
5. conclusioni

TEST DI SIGNIFICATIVITÀ PER $\mu_2 - \mu_1$

1. assunzioni

a) casualità → dati ottenuti utilizzando la randomizzazione

b) distribuzione normale nella popolazione (robusto, soprattutto con n grande)

TEST DI SIGNIFICATIVITÀ PER $\mu_2 - \mu_1$

2. ipotesi

H_o: $\mu_1 = \mu_2$ $(\mu_2 - \mu_1 = 0)$

H_a: $\mu_1 \neq \mu_2$ $(\mu_2 - \mu_1 \neq 0)$ bidirezionale

oppure H_a: $\mu_1 > \mu_2$ $(\mu_2 - \mu_1 < 0)$ unidirezionale

H_a: $\mu_1 < \mu_2$ $(\mu_2 - \mu_1 > 0)$ unidirezionale

TEST DI SIGNIFICATIVITÀ PER $\mu_2 - \mu_1$

3. test

$t = (\bar{y}_2 - \bar{y}_1)/se$

dove $se = \sqrt{(s_1^2/n_1 + s_2^2/n_2}$

TEST DI SIGNIFICATIVITÀ PER $\mu_2 - \mu_1$

4. P-valore

il P-valore è il peso dell'evidenza contro H_o

5. conclusioni

interpretazione del P-valore → piccoli valori di P forniscono forti evidenze contro H_o e a favore di H_a

Si rifiuta H_o se $P \leq \alpha$ (livello di significatività)

NON si rifiuta H_o se $P \geq \alpha$ (livello di significatività)

PROBLEMA B

CONFRONTO TRA 2 MEDIE - campioni dipendenti

primo step:
per ogni coppia di dati calcolare la DIFFERENZA

secondo step:
calcolo la media campionaria delle differenze: $\bar{y}_d$ (che stima μ_d)
e calcolo anche la deviazione standard delle differenze s_d

INTERVALLO DI CONFIDENZA

stima puntuale ± margine d'errore

$$\bar{y}_d \pm t\,(se)$$

$se = s_d / \sqrt{n}$
$gdl = n - 1$

→ stessa formula di quello utilizzato per una singola media

TEST

H_o: $\mu_d = 0$
H_a: $\mu_d \neq 0$ (oppure unidirezionale)

test t per differenze appaiate

$$t = \bar{y}_d / se$$

$se = s_d / \sqrt{n}$ $\quad$ $gdl = n - 1$

→ stessa formula di quella utilizzata per una singola media

ESERCITAZIONE n.7

ESERCIZIO 1

Un sondaggio ha stimato che la percentuale di fumatori adulti in America sia del 41.9% nel 1965 e del 21.5% nel 2003.

a) stima la differenza tra proporzioni di fumatori nei due anni
b) supponi che l'errore standard riportato fosse uguale a 0.020 per ciascuna proporzione; trova l'errore standard della differenza e interpreta

Soluzione

$\prod$-cappello$_1$ = proporzione campionaria di adulti fumatori nel 1965 = 0.419

$\prod$-cappello$_2$ = proporzione campionaria di adulti fumatori nel 2003 = 0.215

a) 0.419 - 0.215 = 0.204
b) se = $\sqrt{0.02^2 + 0.02^2}$ = 0.028
   l'errore standard della differenza descrive con quale precisione il parametro differenza tra proporzioni viene stimato (descrive quanto precisamente 0.204 stima $\prod_1 - \prod_2$

ESERCIZIO 2

Un sondaggio ha indagato la disponibilità dei cittadini UE a pagare di più per avere energia prodotta da fonti rinnovabili; la proporzione di coloro che hanno risposto SI varia da un valore alto registrato in Danimarca pari a 0.52 (n=1008) a uno basso pari a 0.14 registrato in Lituania (n=1002).

a) stima la differenza tra proporzioni di risposte SI delle popolazioni di Danimarca e Lituania
b) le stime delle proporzioni hanno se=0.0157 per la Danimarca e se=0.110 per la Lituania. Usa tali valori per trovare lo se della differenza delle stime di cui al punto (a)

Soluzione

a) 0.52 - 0.14 = 0.204
b) se = $\sqrt{0.0157^2 + 0.110^2}$ = 0.121

ESERCIZIO 3

Per un campione casuale di canadesi, il 60% dichiara di approvare l'operato del primo ministro. Una simile indagine un mese dopo ha una percentuale di favorevoli del 57%. L'intervallo di confidenza al 99% per il cambiamento di opinione nelle proporzioni di popolazione è (-0.07, 0.01): è plausibile che non ci sia stato alcun cambiamento nella percentuale dei sostenitori? perché?

Soluzione

È plausibile che non ci sia stato alcun cambiamento poiché l'intervallo di confidenza include lo 0 e perciò è verosimile che $\prod_2 - \prod_1$ sia uguale a 0

ESERCIZIO 4

Un sondaggio ha esaminato il consumo di alcool tra gli studenti del 2° anno di un'università francese. La percentuale di coloro che hanno dichiarato di bere alcolici più di 4 volte a settimana era pari a 39.9% di 12708 studenti nel 1997 e a 48.2% di 8783 studenti nel 2017.

a) calcola l'errore standard per la stima della differenza tra le proporzioni nel 2017 e nel 1997
b) mostra l'intervallo di confidenza al 95% per la differenza

Soluzione

a) se = $\sqrt{(0.399 \times 0.601)/12708 + (0.482 \times 0.518)/8783} = 0.0069$

b) 95 % IC per $\prod_2 - \prod_1 = 0.083 \pm 0.012$ (0.071, 0.095)

ESERCIZIO 5

La tabella seguente sintetizza le risposte raccolte negli anni 1977 e 2006 alla domanda "è molto meglio per tutti che l'uomo lavori e la donna si occupi della casa e della famiglia". Indica con $\prod_1$ la proporzione di popolazione che concordava con l'affermazione nel 1977 e con $\prod_2$ la stessa proporzione nel 2006.

a) stima la differenza tra proporzioni di risposte nel 2006 e nel 1977 e calcola l'errore standard

b) calcola l'intervallo di confidenza al 95% per $\prod_1 - \prod_2$

c) spiega come i risultati potrebbero differire nel confrontare le proporzioni di chi NON concorda nei due anni

anno	favorevole	NON favorevole	totale
1977	989	514	1503
2006	704	1264	1968

Soluzione

a) 0.658 - 0.358 = 0.30
se = $\sqrt{(0.658 \times 0.342)/1503 + (0.358 \times 0.642)/1968} = 0.0163$

b) 95 % IC per $\prod_1 - \prod_2 = 0.30 \pm 0.03$ (0.27, 0.33)

c) $\prod_3$ = proporzione di popolazione che NON concordava con l'affermazione nel 1977
$\prod_4$ = proporzione di popolazione che NON concordava con l'affermazione nel 2006
la stima ha lo stesso valore assoluto ma cambia il segno e diventa: -0.30
se rimane lo stesso
95 % IC per $\prod_3 - \prod_4 = -0.30 \pm 0.03$ (-0.33, -0.27)

ESERCIZIO 6

Facendo riferimento al problema precedente sul ruolo della donna, nel 2004, 153 su 411 maschi e 166 su 472 femmine hanno risposto SI.

a) definisci la notazione e specifica le ipotesi per verificare l'uguaglianza delle proporzioni di chi ha risposto SI tra i maschi e le femmine

b) stima la proporzione di popolazione presumendo vera H_0, trova l'errore standard della differenza della stima campionaria delle proporzioni e ricava la statistica test

c) ricava il P-valore per l'ipotesi alternativa bidirezionale e interpreta

d) di 652 rispondenti con un titolo di studio inferiore alla laurea, il 40.0% hanno risposto SI. Di 231 soggetti con un titolo di studio superiore o uguale alla laurea, il 25.5% hanno risposto SI. Quale variabile, genere o titolo di studio, sembra influenzare di più l'opinione?

Soluzione

a) $\prod_1$ = proporzione che ha risposto SI tra la popolazione dei maschi
$\prod_2$ = proporzione che ha risposto SI tra la popolazione delle femmine
$\prod$-cappello$_1$ = 153/411 = 0.372
$\prod$-cappello$_2$ = 166/472 = 0.352
H_0: $\prod_1 - \prod_2 = 0$
H_a: $\prod_1 - \prod_2 \neq 0$

b) $\prod$-cappello = stima conglobata = 0.36

$se_0 = \sqrt{(0.36)\,(0.64)\,(1/411 + 1/472)} = 0.0324$

$z = (0.372 - 0.352) / 0.0324 = 0.62$

c) p-valore = 2 x (0.2672) = 0.544

con $\alpha = 0.05$, non rifiuto H_0

d) Π-cappello$_3$ = 0.400

Π-cappello$_4$ = 0.255

H_0: $\Pi_3 - \Pi_4 = 0$

H_a: $\Pi_3 - \Pi_4 \neq 0$

Π-cappello = stima conglobata = 320/883 = 0.36

$se_0 = \sqrt{(0.36)\,(0.64)\,(1/652 + 1/231)} = 0.0368$

$z = (0.400 - 0.255) / 0.0368 = 3.94$

p-valore < 0.001

con $\alpha = 0.05$, rifiuto H_0

La variabile titolo di studio sembra influenzare di più l'opinione

ESERCIZIO 7

Un sondaggio ha rivolto ad alcuni intervistati la seguente domanda "quanti giorni negli ultimi sette ti sei sentito triste?". La media campionaria è di 1.8 per le donne e 1.4 per gli uomini, con un intervallo di confidenza al 95% per il confronto dei due valori pari a (0.2, 0.6) e una statistica test t di 4.8 con p-valore di 0.000. Interpreta i risultati.

Soluzione

L'intervallo di confidenza al 95% per il confronto dei due valori non comprende lo 0, ma solo valori positivi per cui è plausibile che μ_{donne} sia maggiore di μ_{uomini}; tale interpretazione è confermata dal test t e dal p-valore altamente significativo che permette di rifiutare H_0 (con H_0: $\mu_{donne} - \mu_{uomini} = 0$).

ESERCIZIO 8

I risultati di una ricerca che ha confrontato maschi e femmine rispetto al numero di numero di ore al giorno in cui il soggetto guarda la tv hanno fornito:

Group	N	Mean	StDev	SE Mean
F	1117	2.99	2.34	0.070
M	870	2.86	2.22	0.075

a) conduci un test di significatività per analizzare se le medie di popolazione differiscono tra i maschi e le femmine. Riporta le conclusioni per il livello di significatività α=0.05

b) un intervallo di confidenza al 95% per il confronto di medie conterrebbe lo 0?

c) pensi che la distribuzione delle ore dedicate a guardare la tv sia approssimativamente normale? la risposta a questa domanda ha qualche implicazione per le conclusioni tratte nel punto (a)?

Soluzione

a) campioni indipendenti

H_0: $\mu_1 - \mu_2 = 0$

H_a: $\mu_1 - \mu_2 \neq 0$

$se = \sqrt{(0.070)^2 + (0.075)^2} = 0.103$

$t = (2.99 - 2.86)/0.103 = 1.26$

p-valore = 2 x (0.1038) = 0.208

con $\alpha = 0.05$, non rifiuto H_0

b) Sì; 95% IC per $\mu_1 - \mu_2 = 0.13 \pm 0.20$ (-0.07, 0.33)

c) No, perché il valore minimo è a poco più di 1 deviazione standard dalla media; non c'è nessuna implicazione per le conclusioni tratte nel punto (a) perché il test è robusto se n è molto grande (la distribuzione campionaria delle differenze tra le medie campionarie ha forma approssimativamente campanulare con n così grande).

ESERCIZIO 9

Per un'esercitazione una studentessa ha campionato 10 studenti per investigare sulle più comuni attività sociali. In particolare, agli intervistati è stato chiesto di stabilire quante volte al mese nell'anno precedente avevano preso parte alle seguenti attività: andare al cinema e andare ad un evento sportivo. I dati ottenuti sono riportati nella tabella sotto.

a) volendo confrontare le medie tra la frequenza cinema e la frequenza a eventi sportivi facendo inferenza, come dovrebbero essere trattati i campioni, dipendenti o indipendenti?

b) per l'analisi al punto (a) il software mostra i risultati:

	N	Mean	StDev	SE Mean
movies	10	13.000	13.174	4.166
sport	10	9.000	8.380	2.650
difference	10	4.000	16.166	5.112

```
95% IC for mean difference:(-7.56, 15.56)
T-test of mean difference: 0 (vs not=0)
T-value=0.78    p-value=0.454
```

interpreta l'intervallo di confidenza al 95 % mostrato

c) mostra come è stata ottenuta la statistica test t e riporta le conclusioni per il livello di significatività α=0.05

studente	cinema	sport
1	10	5
2	4	0
3	12	20
4	2	6
5	12	2
6	7	8
7	45	12
8	1	25
9	25	0
10	12	12

Soluzione

a) campioni dipendenti

b) è plausibile che $\mu_1 = \mu_2$ perché l'intervallo di confidenza al 95% contiene lo 0

c) test t per differenze appaiate

H_0: $\mu_d = 0$

H_a: $\mu_d \neq 0$

studente	cinema	sport	d
1	10	5	5
2	4	0	4
3	12	20	-8
4	2	6	-4
5	12	2	10
6	7	8	-1
7	45	12	33
8	1	25	-24
9	25	0	25
10	12	12	0

$\bar{y}_d = 4$
$s_d = 16.166$
$se = s_d / \sqrt{n} = 5.113$
$t = 4 / 5.113 = 0.782$
p-valore $> 2 \times (0.1) = 0.2$
con $\alpha = 0.05$, non rifiuto H_0