CORSO DI STATISTICA (parte 2) - ESERCITAZIONE 7

Dott.ssa Antonella Costanzo

a.costanzo@unicas.it

Esercizio 1. Test di indipendenza tra mutabili

In una ricerca di mercato propedeutica al lancio di un nuovo quotidiano si è chiesto a un campione di 2000 individui di età superiore ai 18 anni se essi acquistano o meno un quotidiano ogni giorno. Per studiare le caratteristiche dei lettori si è inoltre chiesto agli intervistati di indicare il titolo di studio posseduto. Si è così ottenuta la seguente distribuzione:

Titolo di Studio	Si	No	Totale
Nessuno	10	190	200
Elementare	90	310	400
Media	150	650	800
Superiore	230	220	450
Laurea	120	30	150
Totale	600	1400	2000

Si può ipotizzare che esista indipendenza tra il titolo di studio e la scelta di acquistare ogni giorno un quotidiano? Si controlli con un opportuno test scegliendo un livello di significatività dell'1%.

Sol.

Per stabilire se esiste indipendenza tra i caratteri oggetto di analisi dobbiamo utilizzare un test statistico basato sul chi-quadro χ.

Il sistema di ipotesi da sottoporre a verifica è il seguente:

H_0: indipendenza tra il titolo di studio e la scelta di acquistare ogni giorno un quotidiano

H_1: associazione tra la il titolo di studio e la scelta di acquistare ogni giorno un quotidiano

Livello di significatività

$\alpha=0.01$

Definizione della statistica test sotto l'ipotesi nulla:

$$\chi^{oss} = \sum_{i=1}^{h}\sum_{j=1}^{k}\frac{(O_{ij}-E_{ij})^2}{E_{ij}} \sim \chi^2_{\alpha;(j-1)(h-1)}$$

dove $E_{ij} = \frac{n_{i.} \times n_{.j}}{n}$

$i = 1,..h$

$j = 1,..k$

Regola di decisione (regione di rifiuto)

Con il livello di significatività $\alpha = 0.01$ e con i= 5 and j = 2 otteniamo

$\chi^2_{4;0.01} = 13.277$.

Quindi, se $\chi^{oss} > \chi^2_{4;0.01} = 13.277$ si rifiuta l'ipotesi nulla

Valore della statistica test sotto l'ipotesi nulla

Tabella teorica sotto l'ipotesi di indipendenza (frequenze teoriche E_{ij})

Titolo di Studio	Si	No	Totale
Nessuno	60	140	200
Elementare	120	280	400
Media	240	560	800
Superiore	135	315	450
Laurea	45	105	150
Totale	600	1400	2000

$$\chi^{oss} = \sum_{i=1}^{h}\sum_{j=1}^{k}\frac{(O_{ij}-E_{ij})^2}{E_{ij}} = \frac{(10-60)^2}{60} + \frac{(190-140)^2}{140} + \cdots = 392.52$$

Il test è significativo all' 1%, infatti si rifiuta l'ipotesi nulla.

Esercizio 2. Test di bontà di adattamento

In 600 lanci di un dado si sono ottenuti i risultati riportati in tabella. Determinare con un livello di significatività del 5% se il dado può considerarsi truccato.

Risultato	1	2	3	4	5	6
Freq. osservate	94	123	88	102	115	78

Sol

Se il dado non è truccato dovremmo osservare una distribuzione uniforme dei risultati

Risultato	1	2	3	4	5	6
Freq. Osservate n_i	94	123	88	102	115	78
Freq. Teoriche $n_i p_i$	100	100	100	100	100	100

Sistema di ipotesi:

H_0: il dado non è truccato (la distribuzione di frequenza osservata è una realizzazione di una v.c. uniforme discreta)

H_1: il dado è truccato (la distribuzione di frequenza osservata NON è una realizzazione di una v.c. uniforme discreta)

Livello di significatività

α=0.05

La *statistica test* corrisponde quindi nel misurare la discrepanza tra le frequenze osservate e quelle teoriche:

$$\chi^{oss} = \sum_{i=1}^{k} \frac{(n_i - n_i p_i)^2}{n_i p_i} \sim \chi^2_{\alpha;(k-1)}$$

dove *k* rappresenta il numero di modalità.

Regola di decisione:

per il livello di significatività fissato il valore critico è in corrispondenza di: $\chi^2_{0.05;(6-1)} = 11.07$

per cui se il valore della statistica test χ^{oss} è maggiore di $\chi^2_{0.05;(6-1)} = 11.07$ si rifiuta l'ipotesi nulla

Il valore della statistica test sotto l'ipotesi nulla è data da:

$$\chi^{oss} = \sum_{i=1}^{k} \frac{(n_i - n_i p_i)^2}{n_i p_i} = \frac{(94-100)^2}{100} + \frac{(123-100)^2}{100} + \frac{(88-100)^2}{100} + \cdots = 14.22$$

Decisione

Siccome $\chi^{oss} > \chi^2_{0.05;(6-1)} = 11.07$ si rifiuta l'ipotesi nulla, quindi rifiuto l'ipotesi di distribuzione uniforme discreta (il dado è truccato)

Esercizio 3. Il modello di regressione

La seguente tabella riporta la valutazione del docente A (su una scala da 1 a 5) e il voto medio che gli studenti si aspettano nell'esame finale, come rilevati nelle schede di valutazione a fine corso dal docente A.

X=Valutazione del docente	2.8	3.7	4.4	3.6	4.7	3.5	4.1	3.2	4.9	4.2	3.8	3.3
Y=Voto atteso	20	23	26	25	24	22	21	19	27	24	26	20

Sapendo che:

$\bar{X} = 3.85$

$\bar{Y} = 23.08$

$\sum (X_i - \bar{X})(Y_i - \bar{Y}) = 13.65$

$\sum (X_i - \bar{X})^2 = 4.35$

$\sum (Y_i - \bar{Y})^2 = 78.917$

$\sqrt{\frac{\sum (Y_i - \hat{Y})^2}{n-2}} = 1.9$

a) Stimare i coefficienti della retta di regressione che spiega il voto atteso in funzione della valutazione attribuita dal docente

b) costruire un intervallo di confidenza al 95% per il coefficiente angolare della retta stimata

c) Verificare l'ipotesi di significatività del modello usando un livello del 5%

Sol.

a) Il modello di regressione lineare: $Y = \beta_0 + \beta_1 X + \varepsilon$

Stima della retta di regressione:

$$Y = b_0 + b_1 X + e$$

I parametri della retta di regressione

Coefficiente angolare: inclinazione della retta di regressione, come varia in media Y a fronte di un incremento unitario della X

$$b_1 = \frac{Codev(X,Y)}{Dev(X)} = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{\bar{x}})}{\sum_{i=1}^{n}(x_i - \bar{\bar{x}})^2} = \frac{13.65}{4.35} = 3.14$$

Nota: è possibile, in alternativa, esprimere il coefficiente angolare della retta di regressione con la seguente:

$$b_1 = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$$

Intercetta della retta di regressione: indica il valore atteso della variabile di risposta Y quando il predittore X assume valore 0.

$$b_0 = \bar{y} - b_1\bar{x} = 23.08 - (3.14) \times 3.85 = 10.99$$

La retta di regressione stimata è pertanto:

$$\hat{Y} = 10.99 + 3.14X$$

b) Per costruire l'intervallo di confidenza al 95% su β_1, abbiamo bisogno di studiare la distribuzione campionaria dello stimatore β_1. Siccome una delle ipotesi classiche del modello di regressione è la normalità degli errori, allora si dimostra che:

$$\beta_1 \sim N\left(\beta_1, \frac{\sigma^2}{\sum_i (x_i - \bar{x})^2}\right)$$

Nota: gli stimatori OLS di β_1 e β_0 sono B.L.U.E.

Tuttavia ciò sarebbe vero (e quindi lo stimatore β_1 si distribuirebbe secondo una legge normale) se conoscessimo la varianza degli errori del modello σ^2. Nella realtà, gli errori del modello non sono osservabili, mentre è possibile osservare i residui. In particolare:

$$RSS = \sum_{i=1}^{n} (Y_i - \hat{Y})^2$$

Lo stimatore non distorto della varianza dei residui del modello che indichiamo con s_e^2 è dato dalla seguente:

$$s_e^2 = \frac{RSS}{n-2}$$

da cui, lo stimatore varianza di b_1

$$s_{b1}^2 = \frac{\frac{RSS}{n-2}}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{s_e^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Distribuzione campionaria dello stimatore b_1 per β_1:

$$\frac{b_1 - \beta_1}{s_{b1}} \sim t_{\frac{\alpha}{2};n-2}$$

dove $s_{b1} = \sqrt{s_{b1}^2}$

Alternativamente, in forma esplicita:

$$\sqrt{\frac{(n-2)\sum_{i=1}^{n}(X_i - \bar{X})^2}{RSS}}(b_1 - \beta_1) \sim t_{\frac{\alpha}{2};n-2}$$

Da cui il corrispondente IC

$$\left[b_1 \pm t_{\frac{\alpha}{2};n-2} s_{b1}\right]$$

oppure, in forma estesa

$$\left[b_1 \pm t_{\frac{\alpha}{2};n-2} \sqrt{\frac{RSS}{(n-2)\sum_{i=1}^{n}(X_i - \bar{X})^2}}\right]$$

Calcoli:

Sapendo che

$$s_e = \sqrt{\frac{\sum(Y_i - \hat{Y})^2}{n-2}} = 1.9$$

A partire da questo risultato si ricava la stima corretta dello scarto quadratico medio del coefficiente di regressione b_1

$$s_{b1} = \frac{s_e}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}} = \frac{1.9}{\sqrt{4.35}} = 0.91$$

$t_{\frac{\alpha}{2};n-2} = t_{0.025;10} = 2.228$

IC per β_1:

$[3.14 \pm 2.228(0.91)]$

c) Un'ipotesi molto importante da verificare nel modello di regressione lineare semplice è che il coefficiente angolare della retta di regressione sia pari a 0: in tal caso, allora la variabile di risposta non dipende dal predittore, in altre parole non c'e regressione sul predittore.

Sistema di ipotesi:

$H_0: \beta_1 = 0$

$H_1: \beta_1 \neq 0$

Livello di significatività $\alpha = 0.05$

Statistica test (sotto H_0): $T = \frac{b_1}{s_{b1}} \sim t_{\frac{\alpha}{2};n-2}$

Regola di decisione

Se $|T| > t_{\frac{\alpha}{2};n-2} = 2.228$ si rifiuta l'ipotesi nulla

Valore della statistica test

$$T = \frac{3.14}{0.91} = 3.45$$

Decisione: T=3.45>2.228, si rifiuta l'ipotesi nulla