STATISTICA (2) – ESERCITAZIONE 7

11.03.2014

Dott.ssa Antonella Costanzo

Esercizio 1. Test di indipendenza tra mutabili

In un'indagine vengono rilevate le informazioni su settore produttivo (Y) e genere (X) su un campione casuale di occupati:

X|Y	**agricoltura**	**artigianato**	**industria**	**servizi**	**totale**
F	0	8	12	80	100
M	10	52	58	20	140
totale	10	60	70	100	240

Testare ad un livello di significatività del 5% se i due caratteri possono essere considerati indipendenti.

Soluzione

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: X e Y sono indipendenti
H_1: X e Y non sono indipendenti

Livello di significatività
α=0.05
Definizione della statistica test sotto l'ipotesi nulla:

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

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

$i = 1,..h$, righe; $j = 1,..k$, colonne

Regola di decisione (regione di rifiuto)

Con il livello di significatività α = 0.05 e con k= 4 e h = 2 otteniamo

$\chi^2_{3;0.05} = 7.815$.

dunque, se $\chi^{stat} > \chi^2_{3;0.05} = 7.815$ si rifiuta l'ipotesi nulla

Calcolo il valore della statistica test sotto l'ipotesi nulla

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

X|Y	**agricoltura**	**artigianato**	**industria**	**servizi**	**totale**
F	4.17	25	29.17	41.67	100
M	5.83	35	40.83	58.33	140
totale	10	60	70	100	240

$$\chi^{stat} = \sum_{i=1}^{h}\sum_{j=1}^{k} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = \frac{(0-4.17)^2}{4.17} + \frac{(10-5.83)^2}{5.83} + \cdots = 104.738$$

Formula alternativa per il calcolo di χ^{stat}

$$\chi^{stat} = n_{..}\left(\sum_{i=1}^{h}\sum_{j=1}^{k} \frac{n_{ij}^2}{n_{i.} \cdot n_{.j}} - 1\right)$$

ovvero:

$$\chi^{stat} = 240\left[\left(\frac{0^2}{100*10} + \frac{8^2}{100*60} + \frac{12^2}{100*70} + . + \frac{10^2}{140*10} + ..\right) - 1\right] = 104.738$$

Decisione:

Essendo $\chi^{stat} > \chi_{0.05,3}$ si rifiuta l'ipotesi nulla al livello di significatività α=5%.

Esercizio 2. Test sulla bontà di accostamento

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

Soluzione

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 $\hat{n}_i = np_i$	100	100	100	100	100	100

dove $p_i = \frac{1}{6}$

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

Costruzione della statistica test

$$\chi^{stat} = \sum_{i=1}^{k} \frac{(n_i - \hat{n}_i)^2}{\hat{n}_i} \sim \chi^2_{\alpha;(k-1)}$$

dove *k* rappresenta il numero di modalità.

Regola di decisione (regione di rifuto):

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^{stat} 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:

$$\chi^{stat} = \frac{(94-100)^2}{100} + \frac{(123-100)^2}{100} + \frac{(88-100)^2}{100} + \cdots = 14.22$$

Decisione

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

Esercizio 3. Test sulla bontà di accostamento (2)

Nell'arco di un triennio, sono stati registrati 1588 incidenti stradali capitati a 706 guidatori di una società di trasporto pubblico. La seguente tabella riporta come tali incidenti sono distribuiti tra i vari autisti:

n. di incidenti	n. di autisti
0	117
1	157
2	158
3	115
4	78
5	44
6	21
7	16

Verificare al un livello di significatività del 5% se questi dati sono compatibili con l'ipotesi che il numero di incidenti per autista abbia una distribuzione di *Poisson*?

Soluzione

Il primo passo nel test di buon adattamento consiste nello stimare il parametro della distribuzione di *Poisson*. Dalla tabella di frequenza si ottiene:

$$\hat{\lambda} = \bar{x} = 2.25$$

Sistema di ipotesi

H_0: i dati seguono una distribuzione di Poisson ($\hat{\lambda} = 2.25$)
H_1: i dati non seguono una distribuzione di Poisson

Livello di significatività
α=0.05

Definizione della statistica test sotto l'ipotesi nulla:

$$\chi^{\text{stat}} = \sum_{\text{i}=1}^{\text{k}} \frac{(n_i - \hat{n}_i)^2}{\hat{n}_i} \sim \chi^2_{\alpha;(\text{k}-1)}$$

dove $\hat{n}_i$= frequenze attese sotto l'ipotesi che i dati seguano una legge di Poisson e
k = numero di modalità

Regola di decisione (Regione di Rifiuto)
Il valore critico in corrispondenza di un livello di significatività del 5% con 8-1=7 gdl è dato da: $\chi_{0.05,7} = 14.07$, per cui se $\chi^{\text{stat}} > \chi_{0.05,7}$ rifiutiamo l'ipotesi nulla.

Calcoliamo le frequenze attese sotto l'ipotesi che il numero di incidenti per autista segua una legge di *Poisson* di parametro $\hat{\lambda}$, ricordando che:

$$X \sim Poi(\lambda);\ \ E(X) = Var(X) = \lambda$$

Distribuzione di probabilità di X:

$$P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!}$$

Es.

$$P(X = 0) = e^{-2.25} = 0.1053 \ \ \rightarrow nP(X = 0) = 706 \cdot 0.1053 = 74.41$$

$$P(X = 1) = \left(\frac{e^{-2.25} \cdot 2.25}{1!}\right) = 0.2371 \ \rightarrow nP(X = 1) = 706 \cdot 0.2371 = 167.42$$

(…)

n. di incidenti	n. di autisti *(freq.osservate)* n_i	n. di autisti *(freq.teoriche sotto H_0)* $\hat{n}_i = n \cdot P(X = x)$
0	117	74.41
1	157	167.42
2	158	188.355
3	115	141.266
4	78	79.46
5	44	35.76
6	21	13.41
7	16	4.31

Il valore della statistica test è pari a:

$$\chi^{\text{stat}} = \frac{(117 - 74.41)^2}{74.41} + \cdots + \frac{(16 - 4.31)^2}{4.31} \approx 57.9$$

Decisione

poiché $\chi^{\text{stat}} > \chi_{0.05,7}$ rifiuto l'ipotesi nulla. I dati non sono compatibili con l'ipotesi che il numero di incidenti per autista segua una legge di *Poisson*.

Esercizio 4. Il modello di regressione: stima, bontà di adattamento, inferenza

Si desidera studiare la relazione tra il voto Y conseguito all'esame di statistica e il voto X conseguito nell'esame di matematica. A partire da un campione casuale di n=200 studenti che hanno sostenuto entrambi gli esami in questione si osservano i seguenti risultati campionari:

$$\frac{1}{200}\sum_{i=1}^{200} y_i = 27.87\,; \frac{1}{200}\sum_{i=1}^{200} x_i = 25.24\,;$$

$$\frac{1}{200}\sum_{i=1}^{200} y_i^2 = 787.52; \frac{1}{200}\sum_{i=1}^{200} x_i^2 = 645.39;$$

$$\frac{1}{200}\sum_{i=1}^{200} x_i\, y_i = 712.51$$

a) Ricavare con il metodo dei minimi quadrati (OLS, *Ordinary Least Squares*), una stima dei parametri del modello di regressione con Y variabile dipendente e X variabile indipendente.
b) Calcolare e interpretare il coefficiente di correlazione lineare e l'indice di determinazione lineare
c) Sulla base del modello stimato, qual è il voto atteso in statistica di uno studente che ha ottenuto un 24 in matematica?
d) Valutare la significatività del modello di regressione (verifica di ipotesi sul coefficiente angolare) con α=0.05.

Soluzione

a) $Y = \beta_0 + \beta_1 X + \varepsilon$ è il modello di regressione lineare per la popolazione

La stima della retta di regressione avviene sulla base del campione:

$Y = \hat{\beta}_0 + \hat{\beta}_1 X + e$

$\hat{\beta}_0$ e $\hat{\beta}_1$ sono rispettivamente gli stimatori per β_0 e β_1 della popolazione

$\hat{\beta}_1$: coefficiente angolare: inclinazione della retta di regressione, come varia in media Y a fronte di un incremento unitario della X

$\hat{\beta}_0$: Intercetta della retta di regressione: indica il valore atteso della variabile di risposta Y quando il predittore X assume valore 0.

Con il metodo dei minimi quadrati, il coefficiente angolare[1] risulta:

$$\hat{\beta}_1 = \frac{Cov(X,Y)}{Var(X)} = \frac{\frac{1}{n}(\sum_{i=1}^{n} x_i\, y_i) - (\frac{1}{n}\sum_{i=1}^{n} y_i)\,(\frac{1}{n}\sum_{i=1}^{n} x_i)}{\frac{1}{n}\sum_{i=1}^{n} x_i^2 - (\frac{1}{n}\sum_{i=1}^{n} x_i)^2}$$

$$\hat{\beta}_1 = \frac{712.51 - 27.87 \cdot 25.24}{645.39 - 25.24^2} = \frac{9.07}{8.33} = 1.09$$

e l'intercetta è pari a:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 27.87 - 1.09(25.24) = 0.36$$

b) Per determinare $\rho_{x,y}$ si può sfruttare la seguente relazione:

$$\rho_{x,y} = \frac{Cov(X,Y)}{\sqrt{Var(X) \cdot Var(Y)}} = \frac{9.07}{\sqrt{8.33 \cdot (787.52 - 27.87^2)}} = \frac{9.07}{\sqrt{8.33 \cdot 10.78}} = 0.96$$

dove $Var(Y) = \sigma_y^2 = \frac{1}{n}\sum_{i=1}^{n} y_i^2 - (\frac{1}{n}\sum_{i=1}^{n} y_i)^2$

Il valore prossimo a 1 di $\rho_{x,y}$ indica la presenza di una forte relazione lineare positiva tra i due voti in questione.

[1] *Nota*: è possibile, in alternativa, esprimere il coefficiente angolare della retta di regressione come:

$$\hat{\beta}_1 = \frac{Codev(X,Y)}{Dev(X)} = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}$$

L'indice di determinazione lineare (R^2) può essere calcolato *anche* come:

$$R^2 = \rho_{x,y}{}^2 = (0.96)^2 = 0.92$$

c) Il modello di regressione stimato è pari a: $\widehat{Y} = 0.36 + 1.09X$ per cui in corrispondenza di $x = 24$ il valore atteso di Y è dato da:

$$\hat{y} = 0.36 + 1.09(24) = 26.52$$

d) Valutare la significatività del modello stimato equivale a testare il seguente

sistema di ipotesi:

$H_0: \beta_1 = 0$

$H_1: \beta_1 \neq 0$

livello di significatività $\alpha = 0.05$

Per costruire la statistica test dobbiamo studiare la distribuzione campionaria di $\hat{\beta}_1$.

Siccome una delle ipotesi del modello di regressione lineare *classico* è

$$\varepsilon_i \sim N(0, \sigma^2) \qquad i.i.d$$

allora si dimostra che[2]:

$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{\sum_i (x_i - \bar{x})^2}\right)$$

Tuttavia ciò sarebbe vero (e quindi lo stimatore per 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:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} (y_i - \hat{y})^2}{n - 2}$$

[2] Nota: gli stimatori OLS di β_1 e β_0 sono *B.L.U.E (Teorema di Gauss-Markov).*

per cui la varianza *corretta* dello stimatore $\hat{\beta}_1$ per β_1 è pari a:

$$S^2_{(\hat{\beta}_1)} = \frac{\hat{\sigma}^2}{\sum_i (x_i - \bar{x})^2} = \frac{\frac{\sum_{i=1}^n (y_i - \hat{y})^2}{n-2}}{\sum_i (x_i - \bar{x})^2}$$

Sfrutto questo risultato per costruire la **statistica test**; infatti sotto H_0 essa risulta:

$$T^{stat} = \frac{\hat{\beta}_1 - \beta_{1|H_o}}{\sqrt{S^2_{(\hat{\beta}_1)}}} \sim t_{\alpha/2;n-2}$$

Nota: per il TLC essendo n sufficientemente grande (n = 200) è ragionevole approssimare la distribuzione di T^{stat} ad una normale standardizzata, in particolare:

$$T^{stat} = \frac{\hat{\beta}_1 - \beta_{1|H_o}}{\sqrt{S^2_{(\hat{\beta}_1)}}} \xrightarrow[n\to\infty]{} N(0,1)$$

Regola di decisione

Con un livello di significatività del 5%, essendo il test bidirezionale e sfruttando l'approssimazione normale (TLC) i valori critici da determinare sono in corrispondenza di: $\pm z_{\alpha/2} = \pm z_{0.975} = \pm 1.96$ per cui:

se $|T^{stat}| > z_{0.975}$ rifiuto l'ipotesi nulla.

Calcoli

Dobbiamo determinare la quantità $S^2_{(\hat{\beta}_1)}$. Sapendo che:

$$\sum_i (x_i - \bar{x})^2 = n \cdot \sigma_x^2 = 200 \cdot 8.33 = 1666$$

E, sfruttando il fatto che:

$R^2 = \rho_{x,y}{}^2 = 0.92$ lo possiamo scrivere[3] come:

[3] Ricorda la definizione dell'indice di determinazione lineare

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

dobbiamo risolvere rispetto alla quantità a numeratore del rapporto (in rosso), per cui:

$$\sum_{i=1}^{n}(y_i - \hat{y})^2 = (1 - R^2)\sum_{i=1}^{n}(y_i - \bar{y})^2$$

dove:

$$\sum_i (y_i - \bar{y})^2 = n \cdot \sigma_y^2 = 200 \cdot 10.78 = 2156$$

quindi:

$$\sum_{i=1}^{n}(y_i - \hat{y})^2 = (1 - 0.92) \cdot 2156 = 172.48$$

Possiamo dunque calcolare:

$$S^2_{(\hat{\beta}_1)} = \frac{\hat{\sigma}^2}{\sum_i (x_i - \bar{x})^2} = \frac{\frac{\sum_{i=1}^{n}(y_i - \hat{y})^2}{n-2}}{\sum_i (x_i - \bar{x})^2} = \frac{\frac{172.48}{200-2}}{1666} = 0.000522$$

E finalmente il **valore della statistica test**:

$$T^{stat} = \frac{1.09 - 0}{\sqrt{0.000522}} = \frac{1.09}{0.0228} = 47.81$$

Decisione

Siccome $|T^{stat}| > z_{0.975}$ rifiuto l'ipotesi nulla.