CORSO DI STATISTICA (parte 2) - ESERCITAZIONE 8

Dott.ssa Antonella Costanzo

a.costanzo@unicas.it

Esercizio 1. Test delle ipotesi sulla varianza

In un'azienda che produce componenti meccaniche, è stato introdotto un nuovo macchinario per la produzione di bulloni. Il diametro dei bulloni prodotti dalla nuova macchina segue una distribuzione normale con media μ e varianza σ^2entrambe incognite. Per valutare la qualità della produzione ottenuta attraverso il nuovo macchinario si misura il diametro di un campione di 4 bulloni prodotti, ottenendo i risultati seguenti 1.8, 2.4, 2.8 3.
Verificare al livello α = 0.01 il sistema di ipotesi seguente:

$H_0: \sigma^2 = 1.5$

$H_1: \sigma^2 > 1.5$

Sol

$\alpha = 0.01$

Dato che la media della popolazione dei diametri dei bulloni prodotti dalla macchina è non nota, la statistica test da utilizzare per verificare il sistema di ipotesi sulla varianza è:

$$C_{stat} = \frac{(n-1)S^2}{\sigma_0^2} \rightarrow \chi^2_{\alpha,n-1}$$

Il livello di significatività è $\alpha = 0.01$, il test è a una coda, quindi

$\chi^2_{0.01,3} = 11.341$

Regola di decisione (regione di rifiuto):

Se $C_{stat} \geq 11..341$ allora rifiutiamo l'ipotesi nulla

Sapendo che $\bar{x} = \frac{1.8+2.4+2.8+3}{4} = 2.5$ e che $S^2 = \frac{0.49+0.01+0.09+0.25}{3} = 0.093$

$$Z_{stat} = \frac{3(0.093)}{1.5} = 0.187$$

Poichè $C_{stat} = 0.187 \leq 11.341$ allora si accetta l'ipotesi nulla.

Esercizio 2. Test dell'indipendenza tra mutabili

Alcuni ricercatori sono interessati a valutare se esiste un'associazione tra l'area di residenza delle famiglie (urbana o rurale) e la presenza di figli minorenni (si o no). A tale proposito viene selezionato un campione casuale di 500 famiglie su cui sono state raccolte le seguenti informazioni:

		Area di residenza	
		Urbana	Rurale
Presenza di figli minorenni	SI	180	145
	NO	80	95

Verificare l'ipotesi di indipendenza tra i due caratteri in tabella un livello di significatività $\alpha = 0.01$.

Al fine di stabilire se esiste o meno associazione tra i caratteri oggetto di studio dobbiamo utilizzare un test statistico basato sul chi-quadro χ.
Il sistema di ipotesi da sottoporre a verifica è il seguente:

H_0: non esiste associazione tra la presenza di figli minorenni e l'area di residenza della famiglia
H_1: esiste associazione tra la presenza di figli minorenni e l'area di residenza delle famiglie

b)
Definizione della statistica test sotto l'ipotesi nulla:

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

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= 2 and j = 2 otteniamo

$\chi^2_{1;0.01} = 6.63$.

Quindi, se C> $\chi^2_{1;0.01} = 6.63$ si rifiuta l'ipotesi nulla e si conclude che esiste associazione tra i caratteri oggetto di studio

Tabella teorica sotto l'ipotesi di indipendenza

		Area di residenza	
		Urbana	Rurale
Presenza di figli minorenni	SI	169	156
	NO	91	84

Per comodità si riportano nella seguente tabella i valori:

$$\frac{\left(O_{ij}-E_{ij}\right)^2}{E_{ij}}$$

		Area di residenza	
		Urbana	Rurale
Presenza di figli minorenni	SI	0.716	0.7756
	NO	1.3297	1.4405

il valore della statistica test è dunque C=4.2618.

Decisione:

C= 4.2618 < 6.63 non si può rifiutare l'ipotesi nulla; non c'è evidenza di un legame associativo tra la presenza di figli minorenni e l'area di residenza della famiglia.

Esercizio 3. Modello di regressione, bontà di adattamento, test sulla significatività dei coefficienti

Le seguenti variabili sono state registrate a partire da un campione casuale di 5 impiegati:

X= tempo di permanenza in ufficio in una settimana lavorativa (in ore)

Y=spesa per cancelleria (in euro)

	x_i	y_i
	35.5	50.6
	27.2	44.1
	30.6	45.9
	35.1	52.6
	38.1	44.8
Totale	166.5	238

1) Valutare, a partire da un modello di regressione lineare se la spesa sostenuta per la cancelleria Y dipende dal tempo di permanenza in ufficio degli impiegati, X. Stimare i parametri della retta di regressione e calcolare il coefficiente di determinazione del modello (bontà di adattamento)

2) Costruire un intervallo di confidenza al 95% per il coefficiente angolare della retta di regressione b_1

3) b_1, misura l'effetto che una variazione unitaria della variabile esplicativa X produce sulla variabile dipendente Y. Sottoporre a verifica delle ipotesi la significatività del coefficiente stimato con il modello di regressione

4) Costruire un test con un livello di significatività del 5% per R^2

Sol.

Tabella dei calcoli

	x_i	y_i	$y_i - \bar{y}$	$x_i - \bar{\bar{x}}$	$(y_i - \bar{y})(x_i - \bar{\bar{x}})$	$(x_i - \bar{\bar{x}})^2$	$(y_i - \bar{\bar{y}})^2$
	35.5	50.6	3	2.2	6.6	4.84	9
	27.2	44.1	-3.5	-6.1	21.35	37.21	12.25
	30.6	45.9	-1.7	-2.7	4.59	7.29	2.89
	35.1	52.6	5	1.8	9	3.24	25
	38.1	44.8	-2.8	4.8	-13.44	23.04	7.84
totale	166.5	238			28.1	75.62	56.98

Dati:

$\bar{x} = 33.3 \qquad \bar{y} = 47.6$

$\sigma_x^2 = 15.124 \qquad \sigma_y^2 = 11.396$

$\sigma_x = 3.89 \qquad \sigma_y = 3.38$

$\sum_{i=1}^{n} (\hat{y} - \bar{y})^2 = 10.44$

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

$$b_1 = \frac{Codev(X,Y)}{Dev(X)} = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{\bar{x}})}{\sum_{i=1}^{n}(x_i - \bar{\bar{x}})^2} = \frac{28.1}{75.62} = 0.3716$$

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.

$b_0 = \bar{y} - b_1 \bar{x} = 47.6 - (0.3716) \times 33.3 = 35.2257$

La retta di regressione stimata è pertanto:

$$\hat{Y} = 35.2257 + 0.3716X$$

Coefficiente di correlazione:

$$\rho_{x,y} = \frac{Cov(X,Y)}{\sigma_x \times \sigma_y} = \frac{28.1/5}{3.89 \times 3.38} = \frac{5.62}{13.148} = 0.43$$

L'obiettivo di un modello di regressione semplice lineare è quello di spiegare come varia la variabile di risposta Y in funzione di una variabile esplicativa X. Il criterio per individuare la retta che meglio descrive la dipendenza funzionale tra le due variabili utilizza questa scomposizione della varianza:

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

Tra le infinite rette che passano per il punto di coordinate $(\bar{x}, \bar{y})$ la retta di regressione è quella che rende minima la devianza residua e, nello stesso tempo, rende massima la devianza di regressione, ovvero:

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

Di conseguenza, tanto maggiore è la variabilità della Y spiegata da X tanto più soddisfacente sarà il modello stimato. Il coefficiente di determinazione (r-quadro) è una misura della bontà di adattamento del modello ai dati, infatti consente di individuare quanta parte della variabilità complessiva di Y è spiegata dalla regressione (vedi esercitazione n.6, prima parte). In particolare:

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

Dove:

$$SST = \sum_{i=1}^{n}(y_i - \bar{y})^2 = 56.98$$

$$SSR = \sum_{i=1}^{n}(\hat{y} - \bar{y})^2 = 10.44$$

$$SSE = \sum_{i=1}^{n}(y_i - \hat{y})^2 = 46.54$$

$$R^2 = \frac{SSR}{SST} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = \frac{10.44}{56.98} = 0.183$$

E' utile inoltre ricordare che si può esprimere il valore del coefficiente di determinazione sfruttando la devianza degli errori:

$$R^2 = 1 - \frac{SSE}{SST}$$

Nella regressione lineare il coefficiente di determinazione può essere ottenuto anche a partire dal coefficiente di correlazione. In particolare, vale la seguente: $R^2 = \rho_{xy}^2 = 0.43^2 = 0.183$.

Nota: in caso di regressione lineare semplice, il coefficiente di correlazione di Pearson può essere espresso come:

$$\rho_{x,y} = b_1 \times \frac{\sigma_x}{\sigma_y} = 0.3716 \times \frac{3.89}{3.38} = 0.43$$

Dunque, conoscendo deviazione standard e coefficiente di regressione possiamo calcolare il coefficiente di Pearson; e viceversa.

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:

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

Per dimostrare questo risultato osserviamo che:

b_1 rappresenta una combinazione lineare delle Y_i, infatti con opportuni passaggi algebrici esso si può esprimere come:

$$b_1 = \frac{x_i - \bar{x}}{\sum_i (x_i - \bar{x})^2} Y_i$$

dove, ovviamente $Y_i \sim N$, i.i.d.

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. A partire da questo, occorre dunque stimare un ulteriore parametro s_e^2 che rappresenta uno stimatore non distorto della varianza dei residui del modello.

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

da cui, lo stimatore varianza di b_1

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

Distribuzione dello stimatore per β_1:

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

Intervallo casuale per β_1 :

$$P\left(b_1 - t_{\frac{\alpha}{2};n-2}\frac{s_e}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}} \leq \beta_1 \leq b_1 + t_{\frac{\alpha}{2};n-2}\frac{s_e}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}}\right) = 1-\alpha$$

Calcoli:

Sapendo che $SSE = \sum_{i=1}^{5}(y_i - \hat{y}_i)^2 = 46.54$ allora

$$s_e^2 = \frac{SSE}{n-2} = \frac{46.54}{3} = 15.513$$

A partire da questo risultato si ricava la stima corretta della varianza del coefficiente di regressione b_1

$$s_{b1}^2 = \frac{s_e^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2} = \frac{15.513}{75.62} = 0.2051$$

da cui:

$$\frac{s_e}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}} = s_{b1} = 0.4529$$

$$t_{\frac{\alpha}{2};n-2} = t_{0.025;3} = 3.182$$

IC per β_1:

$$[-1.069; 1.817]$$

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 $|T| > t_{\frac{\alpha}{2};n-2} = 3.182$ si rifiuta l'ipotesi nulla

Valore della statistica test

$$T = \frac{0.3716}{0.4529} = 0.82$$

Decisione: T=0.82<3.182, non si rifiuta l'ipotesi nulla; il coefficiente stimato non è statisticamente significativo non esiste un legame di dipendenza lineare tra Y e X

d) Test sull'indice di bontà di adattamento del modello ai dati.

Sistema di ipotesi

$H_0: R^2 = 0$

$H_1: R^2 > 0$

Livello di significatività $\alpha = 0.05$

Statistica test (sotto H_0): $F = \frac{R^2(n-2)}{1-R^2} \sim F_{\alpha(num=1;denom=n-2)}$

Regola di decisione:

$F_{\alpha(1;n-1)} = F_{0.05(1;3)} = 10.1$

Se

$F > F_{0.05(1;3)} = 10.1$ allora rifiuto l'ipotesi nulla.

Valore test:

$F = \frac{0.183(5-2)}{1-0.183} = 0.67$

Siccome F=0.67<10.1 non si rifiuta l'ipotesi nulla.