Statistica II modulo (S.I.A.F.A., AA 2020/21)

PREPARAZIONE PER LA PROVA SCRITTA
ESERCIZI COMMENTATI E RISOLTI

Sezione.1: Lettura tavole

Sezione 2: Proprietà stimatori

Sezione 3: Intervalli di confidenza

Sezione 4: Test d’ipotesi

Sezione.5: Regressione lineare semplice

Sezione.1: Lettura tavole

ESERC.1.1.1 (DIFFICOLTÀ: FACILE)

Data una variabile causale normale standardizzata Z, calcolare la probabilità che assuma i valori nel seguente intervallo:

$$\Pr(-2 \leq Z \leq 1) = ?$$

SVOLGIMENTO

Ricorrendo alla tavola che riporta i valori della funzione di ripartizione $\Phi(Z)$della normale standardizzata si può ricavare che la probabilità di $Z \leq 1$ è pari a:

$$\Phi(1) = 0{,}8413$$

Successivamente, applicando le proprietà che caratterizzano la normale standardizzata, si può ricavare che la probabilità di $Z \leq -2$ è pari a :

$$\Phi(-2) = 1 - \Phi(2) = 1 - 0{,}9772 = 0{,}0228$$

per risolvere l'esercizio occorre quindi sottrarre alla probabilità di $Z \leq 1$ la probabilità di $Z \leq -2$:

$$\Pr(-2 \leq Z \leq 1) = \Pr(Z \leq 1) - \Pr(Z \leq -2) = 0{,}8413 - 0{,}0228 = 0{,}8185$$

Quindi si può affermare che la probabilità che una variabile casuale normale standardizzata assuma valori nell'intervallo: -2 – 1 è pari a 81,85%.

RIFERIMENTI PER LO STUDIO

MOODLE (http://e-economiaegiurisprudenza.uniparthenope.it/moodle/course/view.php?id=567)- slides:"Lez1_introduzione_e_cenni_vc"

LIBRO DI TESTO (Borra, Di Ciaccio, Statistica, 4ed.): Cap. 9

Eserc.1.2.2 (**DIFFICOLTÀ: MEDIA)**

Calcolare i valore -Z* e Z* rispetto alla seguente probabilità:

$$\Pr(-z^* \leq Z \leq z^*)=0{,}92$$

SVOLGIMENTO

Per individuare il valore z* si deve considerare che l'intera area al di sotto della funzione di densità è pari a 1 e che l'area tra $-z^*$ e z^* è pari a 0,92. Conseguentemente, il valore z* delimiterà alla sua destra un'area pari a 0,04.

$$\Pr(Z > z^*)=0{,}04$$

Mentre l'area alla sinistra di z*sarà pari a 0,96:

$$\Pr(Z < z^*)=0{,}96$$

Ricorrendo alla tavola dei valori della funzione di ripartizione della Z, il valore z* sarà quindi tale che:

$$\Phi(z^*)=0{,}96$$

Sulla tavola (nel centro della tabella) si individua un'area pari a 0.9599 in corrispondenza del valore di z=1,75. Con una minima approssimazione, si può concludere che z*=1,75 e quindi:

$$\Pr(-z^* \leq Z \leq z^*)=0{,}92 \ = \ \Pr(-1{,}75 \leq Z \leq 1{,}75)=0{,}92$$

Quindi si può affermare che il 92% è la probabilità che una variabile casuale normale standardizzata assuma valori nell'intervallo tra -1,75 e 1,75.

RIFERIMENTI PER LO STUDIO

MOODLE (http://e-economiaegiurisprudenza.uniparthenope.it/moodle/course/view.php?id=567)- slides:"Lez1_introduzione_e_cenni_vc"

LIBRO DI TESTO (Borra, Di Ciaccio, Statistica, 4ed.): Cap. 9

Eserc.1.3.3 (**DIFFICOLTÀ: FACILE)**

Data una variabile T di Student con g=25 ($T_{(25)}$), determinare i valori $T^*_{\frac{a}{2}}$ in base ad un valore di 1-a=0,90:

$$\text{Prob}\ (-t^*_{\frac{a}{2};25} \leq T_{25} \leq t^*_{\frac{a}{2};25}) = \text{1-a} \qquad\qquad (\text{1-a}=0.90)$$

SVOLGIMENTO

Per individuare il valore $\boldsymbol{t^*_{\frac{a}{2};25}}$ si deve considerare che l'intera area al di sotto della funzione di densità è pari a 1 e che l'area tra $-\boldsymbol{t^*_{\frac{a}{2};25}}$ e $\boldsymbol{t^*_{\frac{a}{2};25}}$ è pari a 0,90. Conseguentemente, il valore $\boldsymbol{t^*_{\frac{a}{2};25}}$ delimiterà alla sua destra un'area pari a 0,05 (cioè a/2):

$$\Pr(\text{T} > \boldsymbol{t^*_{\frac{a}{2};25}}) = 0{,}05$$

Ricorrendo alla tavola dei valori critici della T di student, il valore $\boldsymbol{t^*_{0,05;25}}$ si individuerà incrociando i gradi di libertà riportati sulla prima colonna della tavola (pari a 25) con il valore dell'area a destra della T di student riportato nella prima riga (pari a 0,050). Effettuando questo incrocio risulterà un valore $\boldsymbol{t^*_{0,05;25} = 1,7081}$ e quindi:

$$\text{Prob}\ (-t^*_{0,05;25} \leq T_{25} \leq t^*_{0,05;25}) = 0{,}90 = \ \text{Prob}\ (-1{,}7081 \leq T_{25} \leq 1{,}7081\ ) = 0{,}90$$

RIFERIMENTI PER LO STUDIO

MOODLE (http://e-economiaegiurisprudenza.uniparthenope.it/moodle/course/view.php?id=567) - Slides:"Lez.4 Stima intervallare"

LIBRO DI TESTO (Borra, Di Ciaccio, Statistica, 4ed.): Cap. 9

Eserc.1.4.4 (**DIFFICOLTÀ: FACILE)**

Data una variabile Chi-quadrato con g=10 (χ^2_{10}), determinare i valori $\chi^2_{1-\frac{\alpha}{2}}$ e $\chi^2_{\frac{\alpha}{2}}$ in base ad un valore di 1-a=0,99:

$$\text{Prob}\ (\chi^2_{1-\frac{\alpha}{2};10} \leq \chi^2_{10} \leq \chi^2_{\frac{\alpha}{2};10}) = 1\text{-a} \qquad (1\text{-a}=0{,}99)$$

SVOLGIMENTO

Utilizzando le tavole statistiche è possibile individuare i valori della chi-quadrato corrispondenti ad una determinata area sulla coda destra. Nello specifico devono esser ricavati i valori che lasciano a destra un'area a/2 ed un'area pari ad 1-a/2. Considerando che:

1-a=0,99 a=0,01 a/2=0,005 1-a/2=0,995

Dato un valore dei gradi di libertà (g) pari a 10, si verificherà che

$$\chi^2_{0,995;10}=2{,}1558 \qquad \chi^2_{0,005;10}=25{,}1881$$

Quindi:

$$\text{Prob}\ (2{,}1558 \leq \chi^2_{10} \leq 25{,}1881) = 0{,}99$$

RIFERIMENTI PER LO STUDIO

MOODLE (http://e-economiaegiurisprudenza.uniparthenope.it/moodle/course/view.php?id=567) - Slides:"Lez.4 Stima intervallare"

LIBRO DI TESTO (Borra, Di Ciaccio, Statistica, 4ed.): Cap. 9

Eserc.1.5.5 (**DIFFICOLTÀ: FACILE)**

Data una variabile F di Fisher con g_1=11 e g_2=15 ($\boldsymbol{F_{11;15}}$), determinare i valori $F_{1-\frac{\alpha}{2}}$ e $F_{\frac{\alpha}{2}}$ in base alla seguente probabilità:

$$\text{Prob } (f_{1-\frac{\alpha}{2}} \leq F_{11;15} \leq f_{\frac{\alpha}{2}} \ )= 1\text{-a} \qquad\qquad (1\text{-a}=0{,}95)$$

SVOLGIMENTO

Considerando che:

1-a=0,95 a=0,05 a/2=0,025 1-a/2=0,975

Quindi:

$$\text{Prob } (f_{0,975} \leq F_{11;15} \leq f_{0,025})= 0{,}95$$

Il valore $f_{0,025}$ si ricava direttamente dalle tavole statistiche della F di Fisher che riportano i valori della F di Fisher che delimitano una determinata area sulla coda destra della distribuzione; considerando che g_1=11 e g_2=15 si ottiene che

$$f_{0,025;11,15} = 3{,}01$$

Per individuare il valore della F di Fisher che delimita a destra un'area pari a 1-a/2 si utilizza la seguente proprietà che caratterizza la v.c. F di Fisher:

$$f_{0,925;11,15} = \frac{1}{f_{0,025;15,11}} = \frac{1}{3{,}33} = 0{,}30$$

Quindi:

$$\text{Prob } (0{,}30 \leq F_{11;15} \leq 3{,}33 \ \ )= 0{,}95$$

RIFERIMENTI PER LO STUDIO

MOODLE (http://e-economiaegiurisprudenza.uniparthenope.it/moodle/course/view.php?id=567)- slides:"Lez.6 Campioni indipendenti"

LIBRO DI TESTO (Borra, Di Ciaccio, Statistica, 4ed.): Cap. 14

Sezione 2: Proprietà stimatori

Eserc.2.1.6 (DIFFICOLTÀ: FACILE)

Lo stimatore ha una sua distribuzione campionaria. Vero o Falso?

SVOLGIMENTO

Vero. Lo stimatore, essendo funzione delle osservazioni campionarie (dipende dal campione), ha una sua distribuzione campionaria la cui conoscenza permette di capire se lo stimatore scelto produrrà con elevata probabilità stime "vicine" al valore vero del parametro.

Eserc.2.2.7(DIFFICOLTÀ: MEDIA)

L'errore quadratico medio di uno stimatore corretto coincide con la sua varianza. Vero o Falso?

SVOLGIMENTO

Vero. Uno stimatore corretto ha distorsione nulla, quindi $\boldsymbol{MSE(T) = Var(T) + B(T)^2 = Var(T) + 0 = Var(T)}$

Eserc.2.3.8 (DIFFICOLTÀ: FACILE)

Uno stimatore si dice distorto se non è efficiente. Vero o Falso?

SVOLGIMENTO

Falso. Uno stimatore di un parametro θ è distorto se il suolo valore atteso è diverso da θ per qualche valore di θ.

Eserc.2.4.9 (DIFFICOLTÀ: MEDIA)

La distorsione di uno stimatore è sempre positiva. Vero o Falso?

SVOLGIMENTO

Falso. La distorsione può essere sia positiva (lo stimatore sovrastima il vero valore del parametro) che negativa (lo stimatore sottostima il vero valore del parametro).

Eserc.2.5.10 (DIFFICOLTÀ: MEDIA)

Per confrontare due stimatori di uno stesso parametro è sufficiente confrontare le loro varianze. Vero o Falso?

SVOLGIMENTO

Falso Per confrontare due stimatori usando il MSE come criterio è sufficiente confrontare le loro varianze solo se sono corretti.

RIFERIMENTI PER LO STUDIO

MOODLE (http://e-economiaegiurisprudenza.uniparthenope.it/moodle/course/view.php?id=567)- slides:"Lez3_stima_puntuale"

LIBRO DI TESTO (Borra, Di Ciaccio, Statistica, 4ed.): Cap. 11

Sezione 3: Intervalli di confidenza

<u>Eserc.3.1.11</u> (DIFFICOLTÀ: MEDIA)

Un'agenzia immobiliare vuole stimare il prezzo medio di vendita degli appartamenti di una zona di Roma. Considera un campione di 25 vendite e calcola il prezzo medio ottenendo 148.000 euro. Sapendo che la distribuzione del prezzo di vendita degli appartamenti è un v.c. Normale con deviazione standard nota uguale a 22.000 euro, calcolare l'intervallo di confidenza al 95% per il prezzo medio delle vendite.

<u>SVOLGIMENTO</u>

Poiché è nota la varianza (deviazione standard) della popolazione, la formula per il calcolo dell'intervallo di confidenza della media è data da:

$$P\left(\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1 - a$$

È necessario calcolare il valore $z_{\alpha/2}$, considerando che 1-a=0,95 e quindi a/2=0,025, si avrà:

$$z_{0,025} = 1,96$$

Applicando i valori relativi alla media campionaria, all'ampiezza del campione ed alla deviazione standard si ottiene:

$$P\left(148000 - 1,96\frac{22000}{\sqrt{25}} \leq \mu \leq 148000 + 1,96\frac{22000}{\sqrt{25}}\right) = 0,95$$

$$P(148000 - 8624 \leq \mu \leq 148000 + 8624) = 0,95$$

Infine:

$$P(139376 \leq \mu \leq 156624) = 0,95$$

Si può quindi affermare che, con una probabilità del 95%, la media dei prezzi di vendita degli appartamenti sarà compresa tra 139376 euro e 156624 euro.

<u>RIFERIMENTI PER LO STUDIO</u>

MOODLE (http://e-economiaegiurisprudenza.uniparthenope.it/moodle/course/view.php?id=567)- slides:"Lez.4 Stima intervallare"

LIBRO DI TESTO (Borra, Di Ciaccio, Statistica, 4ed.): Cap. 12

<u>Eserc.3.2.12</u> (DIFFICOLTÀ: AVANZATA)

Ad un campione di 50 individui residenti in una città è stato chiesto di dare un voto da 1 (pessimo) a 10 (ottimo) al servizio di raccolta rifiuti. La media campionaria del punteggio è pari a 6,5 mentre S= 1,5.

a) Calcolare l'intervallo di confidenza al 95% per il punteggio medio dato dai cittadini.
b) L'intervallo di confidenza al 99% è più ampio o più ristretto rispetto all'intervallo di confidenza al 95%? Perché?
c) Si supponga che la stessa media campionaria x e la stessa deviazione standard siano state osservate su un campione di 100 cittadini. Calcolare il nuovo intervallo di confidenza al 95%. È più o meno ampio dell'intervallo di cui ai punti a. e b.? Perché?

<u>SVOLGIMENTO</u>

Punto a

Per lo svolgimento dell'esercizio si deve considerare che non si conosce la varianza della popolazione ma è stata fornita la deviazione standard campionaria, conseguentemente si utilizzerà la formula dell'intervallo di confidenza della media con varianza incognita uguale a :

$$P\left(\bar{X} - t_{\alpha/2,\mathrm{g}}\frac{S}{\sqrt{n}} \leq \mu \leq \bar{X} + t_{\alpha/2,\mathrm{g}}\frac{S}{\sqrt{n}}\right) = 1 - a$$

È necessario calcolare il valore $t_{\alpha/2,g}$, considerando che 1-a=0,95 (quindi a/2=0,025) e g=n-1=49, si avrà:

$$t_{0,025,49} = 2{,}0096$$

Applicando i valori relativi alla media campionaria, all'ampiezza del campione ed alla deviazione standard si ottiene:

$$P\left(6{,}5 - 2{,}0096\frac{1{,}5}{\sqrt{50}} \leq \mu \leq 6{,}5 + 2{,}0096\frac{1{,}5}{\sqrt{50}}\right) = 0{,}95$$

$$P(6{,}5 - 0{,}4263 \leq \mu \leq 6{,}5 + 0{,}4263) = 0{,}95$$

Infine:

$$P(6{,}0737 \leq \mu \leq 6{,}9263) = 0{,}95$$

Si può quindi affermare che, con una probabilità del 95%, la media del punteggio di soddisfazione dei cittadini per il servizio di raccolta dei rifiuti sarà compresa tra 6,0737 e 6,9263.

Punto b

Se si considera un livello di confidenza pari a 0,99 (invece di 0,95), si assisterà ad un aumento dell'ampiezza dell'intervallo di confidenza in quanto all'aumento del livello di confidenza corrisponde una diminuzione della probabilità "a" e quindi un aumento del valore $t_{\alpha/2,g}$ che nell'esercizio sarà pari a :

$$t_{0,005,49} = 2{,}68$$

Quindi il nuovo intervallo di confidenza sarà pari a:

$$P(5{,}9314 \leq \mu \leq 7{,}0685) = 0{,}95$$

il nuovo intervallo sarà quindi meno accurato (ampiezza maggiore) ma sarà più affidabile (minore valore di “a”).

Punto c

Un aumento dell’ampiezza campionaria permette di diminuire l’ampiezza dell’intervallo, algebricamente l’aumento di n comporta una diminuzione dello std. error ($\frac{S}{\sqrt{n}}$) e quindi una diminuzione della quantità che si aggiunge e sottrae alla media campionaria.

Da un punto di vista teorico, l’aumento dell’ampiezza campionaria si traduce in una diminuzione della varianza dello stimatore della media che conseguentemente produrrà delle stime campionaria più vicine al suo valore atteso (la media della popolazione). Il “nuovo” intervallo di confidenza sarà più accurato e, considerando i dati dell’esercizio con un livello di confidenza del 95% , sarà dato da :

$$P(6{,}5 - 0{,}3014 \leq \mu \leq 6{,}5 + 0{,}3014) = 0{,}95$$

Infine:

$$P(6{,}1985 \leq \mu \leq 6{,}8014) = 0{,}95$$

<u>RIFERIMENTI PER LO STUDIO</u>

MOODLE (http://e-economiaegiurisprudenza.uniparthenope.it/moodle/course/view.php?id=567)- slides:“Lez.4 Stima intervallare”

LIBRO DI TESTO (Borra, Di Ciaccio, Statistica, 4ed.): Cap. 12

Eserc.3.3.13 (DIFFICOLTÀ: MEDIA)

Una famiglia cerca una casa di due stanze da prendere in affitto in una certa zona di una città. Per avere un'idea dei prezzi degli affitti mensili nella zona d'interesse decide di guardare i prezzi riportati nella vetrina di un'agenzia immobiliare.

La tabella in basso riporta i prezzi degli affitti nella zona considerata per abitazioni con due stanze:

775	900	850	900	1100	1000	980	920	880	950

Supponendo che il prezzo degli affitti delle case di medesima tipologia si distribuisca secondo una Normale, calcolare un intervallo di confidenza per il prezzo medio di affitto al 95%.

SVOLGIMENTO

Per poter calcolare l'intervallo di confidenza richiesto, è necessario calcolare la media e la deviazione standard del campione osservato:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{775 + 900 + 850 + \cdots .. + 880 + 950}{10} = \frac{9255}{10} = 925,5$$

$$S^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} = \frac{71322.5}{9} = 7924.72 \qquad S = \sqrt{S^2} = \sqrt{925.5} = 89$$

Data la formula dell'intervallo di confidenza in caso di varianza della popolazione incognita (quindi stimata su base campionaria):

$$P\left[\bar{X} - t_{\alpha/2,n-1}\frac{S}{\sqrt{n}} \leq \mu \leq \bar{X} + t_{\alpha/2,n-1}\frac{S}{\sqrt{n}}\right] = 1 - \alpha$$

È necessario calcolare il valore della t di student ($t_{\alpha/2,n-1}$), considerando che i gradi di libertà sono pari a 9 (10-1) e che 1-a=0,95, si avrà:

$t_{\alpha/2,n-1} = t_{0,025,9}$ =2,2622

Conseguentemente:

$$P\left[925,5 - 2,2622\frac{89}{\sqrt{10}} \leq \mu \leq 925,5 + 2,2622\ \frac{89}{\sqrt{10}}\right] = 0,95$$

$$P\left[925,5 - 2,2622\frac{89}{\sqrt{10}} \leq \mu \leq 925,5 + 2,2622\ \frac{89}{\sqrt{10}}\right] = 0,95$$

$$P[925,5 - 63,66 \leq \mu \leq 925,5 + 63,66] = 0,95$$

$$P[861,83 \leq \mu \leq 989,17] = 0,95$$

In conclusione, si può affermare che, con una probabilità del 95%, la media dei prezzi degli affitti sarà compresa tra 861,83 euro e 989,17 euro.

RIFERIMENTI PER LO STUDIO

MOODLE (http://e-economiaegiurisprudenza.uniparthenope.it/moodle/course/view.php?id=567)- slides:“Lez.4 Stima intervallare”

LIBRO DI TESTO (Borra, Di Ciaccio, Statistica, 4ed.): Cap. 12

Eserc.3.4.14 (DIFFICOLTÀ: MEDIA)

Si vuole stimare la proporzione di individui che utilizzano un personal computer. Vengono intervistate casualmente 370 persone, delle quali 214 affermano di utilizzare il pc. Costruire un intervallo di confidenza, ad un livello di fiducia del 95%, della proporzione di utilizzatori di pc.

SVOLGIMENTO

La formula per il calcolo dell'intervallo di confidenza di una proporzione è data da:

$$P\left(\overline{X} - z_{\frac{\alpha}{2}}\sqrt{\frac{\overline{X}(1-\overline{X})}{n}} < \pi < \overline{X} + z_{\frac{\alpha}{2}}\sqrt{\frac{\overline{X}(1-\overline{X})}{n}}\right) = 1-\alpha$$

È necessario calcolare il valore $z_{\alpha/2}$, considerando che 1-a=0,95 e quindi a/2=0,025, si avrà:

$$z_{0,025} = 1{,}96$$

Applicando i valori relativi alla proporzione campionaria, all'ampiezza del campione ed al valore $z_{\alpha/2}$, si ottiene:

$$P\left(0{,}58 - 1{,}96\sqrt{\frac{0{,}58(1-0{,}58)}{370}} < \pi < 0{,}58 + 1{,}96\sqrt{\frac{0{,}58(1-0{,}58)}{370}}\right) = 0{,}95$$

$$(0{,}53 < \pi < 0{,}63) = 0{,}95$$

Quindi si può affermare, con un livello di confidenza del 95%, che la proporzione di individui che utilizza il computer è compresa tra 0,53 e 0,63 (in termini percentuali tra il 53% ed il 63%).

RIFERIMENTI PER LO STUDIO

MOODLE (http://e-economiaegiurisprudenza.uniparthenope.it/moodle/course/view.php?id=567)- slides:"Lez.4 Stima intervallare"

LIBRO DI TESTO (Borra, Di Ciaccio, Statistica, 4ed.): Cap. 12

Eserc.3.5.15 (**DIFFICOLTÀ: MEDIA)**

Una ditta produce guide turistiche, estraendo un campione casuale di 20 guide si è osservato che l'altezza media di ogni volume è di 28 cm. con una varianza (S^2) di 0,25 cm. Calcolare l'intervallo di confidenza della varianza dell'altezza dei volumi considerando un a=0.05.

SVOLGIMENTO

Per costruire un intervallo di confidenza per la varianza si utilizza la varianza campionaria corretta S^2 come stimatore della varianza della popolazione. Si dimostra che:

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$$

Di conseguenza considerando che

$$Pr\left(\chi^2_{1-\frac{a}{2};\,n-1} \leq \chi^2_{n-1} \leq \chi^2_{\frac{a}{2};\,n-1}\right) = 1-a$$

Si può ricavare, con semplici passaggi matematici, la seguente formulazione dell'intervallo di confidenza:

$$\left(\frac{(n-1)S^2}{\chi^2_{\alpha/2}} < \sigma^2 < \frac{(n-1)S^2}{\chi^2_{1-\alpha/2}}\right) = 1-\alpha$$

Per calcolare l'intervallo è quindi necessario calcolare i valori $\chi^2_{1-\alpha/2}$ e $\chi^2_{\alpha/2}$. Considerando un livello 1-a= 95% ed un valore di g=19, si avrà:

$\chi^2_{0,975;19}$.= 8,9065 $\chi^2_{0,025;19}$ = 32,8523

$$\left(\frac{19 \cdot 0,25}{32,8523} < \sigma^2 < \frac{19 \cdot 0,25}{8,9065}\right) = 0,95$$

$$P(0,1446 < \sigma^2 < 0,5333) = 0,95$$

Quindi si può affermare, con un livello di confidenza del 95%, che la varianza dell'altezza dei volumi sarà compresa tra 0,1446 e 0,5333.

RIFERIMENTI PER LO STUDIO

MOODLE (http://e-economiaegiurisprudenza.uniparthenope.it/moodle/course/view.php?id=567)- slides:"Lez.4 Stima intervallare"

LIBRO DI TESTO (Borra, Di Ciaccio, Statistica, 4ed.): Cap. 12

Eserc.3.6.16 (**DIFFICOLTÀ: MEDIA)**

Si intende costruire un intervallo di confidenza al 95% con ampiezza massima di 4 (semi-lunghezza $\boldsymbol{\delta}$ =2) per la media di una popolazione. Ipotizzando che la varianza della popolazione sia nota e pari a 100, determinare la dimensione campionaria minima.

SVOLGIMENTO

Si deve determinare l'ampiezza campionaria necessaria per ottenere un intervallo di confidenza con un'ampiezza max di 4, i dati disponibili sono i seguenti:

$\boldsymbol{n}$	$\boldsymbol{\delta}$	$\boldsymbol{\sigma^2}$	$\boldsymbol{z_{\alpha/2}}$
?	2	100	$z_{\alpha/2} = z_{0,025} = 1,96$

Per determinare il valore di n si applicherà la seguente formula:

$$n = \left(z_{\alpha/2}\frac{\sigma}{\delta}\right)^2 = \left(1,96\,\frac{10}{2}\right)^2 = 96.04$$

Sarà quindi necessario considerare un campione di numerosità n=96

RIFERIMENTI PER LO STUDIO

MOODLE (http://e-economiaegiurisprudenza.uniparthenope.it/moodle/course/view.php?id=567)- slides:"Lez.4 Stima intervallare"

LIBRO DI TESTO (Borra, Di Ciaccio, Statistica, 4ed.): Cap. 12

Sezione 4: TEST D'IPOTESI

Eserc.4.1.17 (**DIFFICOLTÀ: AVANZATA)**

Su un campione di n = 8 call center si sono rilevate le chiamate ricevute in un giorno, che sono riportate nella tabella seguente:

8	9	6	9	10	7	4	11

Assumendo che il campione sia estratto da una popolazione normale con varianza nota pari a 25:

a) Si verifichi l'ipotesi che la media delle chiamate ricevute giornalmente da tutti i call center sia pari a 6 contro l'alternativa che sia superiore a 6, ad un livello di significatività α = 0,06.

b) Stabilire il livello di α tale che venga ribaltata la decisione effettuata nel test precedente e motivare il cambiamento dal punto di vista teorico.

__SVOLGIMENTO__

Punto a

si dovrà svolgere un test d'ipotesi della media conoscendo l'ammontare della varianza della popolazione, il test avrà il seguente sistema d'ipotesi:

$$\mathrm{H_o}: \mu = 6$$

$$\mathrm{H_1}: \mu > 6$$

si utilizzerà la statistica test Z:

$$Z = \frac{\overline{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$$

Con la seguente regola di decisione:

$$Z \leq z_a \; Si\ accetta\ H_o$$
$$\mathrm{Z > z_a \text{ Si rifiuta } H_o}$$

In base ai dati campionari si ottiene una media campionaria $\overline{X} = 8$ e quindi la statistica test sarà pari a:

$$Z = \frac{8-6}{\frac{5}{\sqrt{8}}} = \frac{2}{\frac{5}{\sqrt{8}}} = 1{,}1313$$

Poiché a=0,06, si avrà:

$$z_a = z_{0,06} = 1{,}555$$

Essendo:

$$Z < z_a \rightarrow Si\ accetta\ H_o$$

Si conclude affermando che, con un livello di confidenza del 94%, si accetta l'ipotesi che la media delle chiamate ricevute giornalmente da tutti i call center sia pari a 6

Punto b

Il livello minimo di a per “ribaltare” la decisione del test sarà quel valore che individuerà una regione di rifiuto in grado di includere anche il valore della statistica test (pari a 1,1313).

Per risolvere l’esercizio basterà ricorrere alle tavole della Z, si osserva che considerando un valore

$$z_a = 1{,}12$$

Si rifiuterà l’ipotesi nulla, in quanto Z=1,1313 e quindi:

$$Z > z_a \rightarrow 1{,}1313 > 1{,}12 \ \rightarrow Si\ rifiuta\ H_o$$

Il valore di a che determina un valore di z_a=1,12 sarà:

a= 1-0,8686=0,1314 $\rightarrow$ $z_{0,1314} = 1{,}12$

In conclusione, per valori di a>=0,1314 si rifiuterà l’ipotesi nulla mentre per valori inferiori si accetterà H_0

<u>RIFERIMENTI PER LO STUDIO</u>

MOODLE (http://e-economiaegiurisprudenza.uniparthenope.it/moodle/course/view.php?id=567) - slides:“Lez.5 Test ipotesi”

LIBRO DI TESTO (Borra, Di Ciaccio, Statistica, 4ed.): Cap. 14

Eserc.4.2.18 (**DIFFICOLTÀ: MEDIA)**

Un società di distribuzione ha rilevato il fatturato di un campione dei suoi punti vendita. Il campione è composto da 9 punti vendita che hanno comunicato i seguenti dati:

Punto vendita	1	2	3	4	5	6	7	8	9
Fatturato (migliaia di euro)	80	95	70	65	100	40	55	75	50

Il target fissato per l'anno 2018 è pari a 60 (mila euro). Si può ritenere che la media del fatturato sia significativamente superiore (con un a=5%) a quella prevista come target?

SVOLGIMENTO

Si dovrà svolgere un test sulla media del fatturato considerando che la varianza della popolazione è incognita. Il sistema di ipotesi sarà:

$$\mathrm{H_o}: \mu = 60$$

$$\mathrm{H_1}: \mu > 60$$

Per svolgere il test, si utilizzerà la statistica test:

$$t = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} \sim t - Student(n-1)$$

Con la seguente regola di decisione:

$$\mathrm{T} \leq \mathrm{t_a} \text{ Si accetta } \mathrm{H_o}$$
$$\mathrm{T} > \mathrm{t_a} \text{ Si rifiuta } \mathrm{H_o}$$

Una volta stimata la media campionaria pari a :

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{630}{9} = 70$$

Si calcolerà la varianza campionaria corretta in base alla seguente formulazione

$$S^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} = \frac{3200}{8} = 400 \qquad S = \sqrt{S^2} = \sqrt{400} = 20$$

Si può quindi calcolare la statistica test:

$$t = \frac{70 - 60}{20/\sqrt{9}} = 1{,}5$$

Poiché a=0,05 e g=9-1=8, si avrà:

$$t_{a,n-1} = t_{0{,}05{,}8} = 1{,}8595$$

Si verifica che:

$$T < t_{a,n-1} \rightarrow Si\ accetta\ H_o$$

Si conclude affermando che, con un livello di confidenza del 95%, si accetta l'ipotesi che la media del fatturato è pari a 60mila euro.

RIFERIMENTI PER LO STUDIO

MOODLE (http://e-economiaegiurisprudenza.uniparthenope.it/moodle/course/view.php?id=567)- slides:"Lez.5 Test ipotesi"

LIBRO DI TESTO (Borra, Di Ciaccio, Statistica, 4ed.): Cap. 14

Eserc.4.3.19 (**DIFFICOLTÀ: MEDIA)**

Mediante indagini di mercato un'azienda ha rilevato che la quota di mercato di un suo prodotto è del 25%. Dopo aver effettuato una campagna pubblicitaria vuole sapere se vi è stato un aumento delle vendite. Da un campione (n=1000) si rileva che il 31% delle persone intervistate acquista il suo prodotto. Verificare, al livello di significatività del 1% se la campagna pubblicitaria è stata efficace nell'incrementare la propria quota di mercato.

SVOLGIMENTO

si dovrà svolgere un test d'ipotesi sulla proporzione (quota di mercato), il test avrà il seguente sistema d'ipotesi:

$$\mathrm{H_o}: \pi = 0{,}25$$

$$\mathrm{H_1}: \pi > 0{,}25$$

si utilizzerà la statistica test Z:

$$Z = \frac{\overline{X} - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}}$$

Con la seguente regola di decisione:

$$Z \leq z_a \; Si\ accetta\ H_o$$
$$\mathrm{Z > z_a \text{ Si rifiuta } H_o}$$

In base ai dati dell'esercizio, si osserva una proporzione campionaria $\overline{X} = 0{,}31$ e quindi la statistica test sarà pari a:

$$Z = \frac{0{,}31 - 0{,}25}{\sqrt{\frac{0{,}25(1-0{,}25)}{1000}}} = \frac{0{,}06}{0{,}0137} = 4{,}379$$

Poiché a=0,01, si avrà:

$$z_a = z_{0,01} = 2{,}326$$

Essendo:

$$Z > z_a \rightarrow Si\ rifiuta\ H_o$$

Si conclude affermando che, con un livello di confidenza del 99%, si rifiuta l'ipotesi che la quota di mercato sia pari al 25% in favore dell'ipotesi che prevede una quota di mercato più elevata.

RIFERIMENTI PER LO STUDIO

MOODLE (http://e-economiaegiurisprudenza.uniparthenope.it/moodle/course/view.php?id=567)- slides:"Lez.5 Test ipotesi"

LIBRO DI TESTO (Borra, Di Ciaccio, Statistica, 4ed.): Cap. 14

Eserc.4.4.20(DIFFICOLTÀ: MEDIA)

Si ipotizzi che la spesa delle famiglie per vacanze si distribuisca secondo una variabile casuale Normale. Estratto un campione di 61 famiglie si osserva una varianza campionaria corretta pari a 90000. Verificare con un a= 0.05 se la varianza della popolazione sia uguale oppure diversa da 72000.

SVOLGIMENTO

si dovrà svolgere un test d'ipotesi sulla varianza, il test avrà il seguente sistema d'ipotesi:

$$\mathrm{H_o}: \sigma^2 = 72000$$

$$\mathrm{H_1}: \sigma^2 \neq 72000$$

si utilizzerà la statistica test χ^2:

$$\chi^2 = \frac{(n-1)S^2}{\sigma_0^2}$$

Considerando che l'ipotesi alternativa è tipo bidirezionale, si utilizzerà la seguente regola di decisione:

$$\begin{array}{ll} \chi^2_{1-\frac{a}{2};g} \leq \chi^2 \leq \chi^2_{\frac{a}{2};g} & \text{si accetta } \mathrm{H_o} \\ \chi^2 > \chi^2_{\frac{a}{2};g} & \text{si rifiuta } \mathrm{H_o} \\ \chi^2 < \chi^2_{1-\frac{a}{2};g} & \text{si rifiuta } \mathrm{H_o} \end{array}$$

In base ai dati dell'esercizio, si osserva una varianza campionaria $S^2 = 90000$ ed una statistica test pari a:

$$Z = \frac{(61-1)\ 90000}{72000} = 75$$

Poiché a=0,05 e g=61-1=60, si avrà:

$$\chi^2_{\frac{a}{2};g} = \chi^2_{0,025;60} = 83{,}2977 \qquad \chi^2_{1-\frac{a}{2};g} = \chi^2_{0,975;60} = 40{,}4817$$

Essendo:

$$\chi^2_{1-\frac{a}{2};g} \leq \chi^2 \leq \chi^2_{\frac{a}{2};g} \quad \rightarrow \quad 40{,}4817 \leq 75 \leq 83{,}2977 \quad \rightarrow \quad \text{si accetta } \mathrm{H_o}$$

Si conclude affermando che, con un livello di confidenza del 95%, si accetta l'ipotesi che la varianza della spesa delle famiglie per vacanze sia pari a 72000.

RIFERIMENTI PER LO STUDIO

MOODLE (http://e-economiaegiurisprudenza.uniparthenope.it/moodle/course/view.php?id=567)- slides:"Lez.5 Test ipotesi"

LIBRO DI TESTO (Borra, Di Ciaccio, Statistica, 4ed.): Cap. 14

Eserc.4.5.21 (**DIFFICOLTÀ: AVANZATA)**

In due città italiane si è misurata la spesa media mensile delle famiglie per attività culturali su due campioni di numerosità n_1=51 e n_2=41, che è risultata pari, rispettivamente, a 104,3 euro e 97,3 euro con varianze campionarie corrette S_1^2=100 e S_2^2=90. Si verifichi al livello del 98% se è possibile affermare che non esiste una differenza nella media della spesa per attività ricreative fra le due città considerate ipotizzando che le varianze delle popolazioni siano uguali.

SVOLGIMENTO

Si dovrà svolgere un test d'ipotesi sulla differenza tra le medie di due popolazioni considerando che le rispettive varianze sono incognite ma uguali, il test avrà il seguente sistema d'ipotesi:

$$\mathrm{H_o}: \mu_1 = \mu_2 \qquad \rightarrow \qquad \mathrm{H_o}: \mu_1 - \mu_2 = 0$$

$$\mathrm{H_1}: \mu_1 \neq \mu_2 \qquad \rightarrow \qquad \mathrm{H_1}: \mu_1 - \mu_2 \neq 0$$

si utilizzerà la statistica test t:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\sigma_p^2(\frac{1}{n_1} + \frac{1}{n_2})}} \sim t - Student(n_1 + n_2 - 2)$$

Considerando che l'ipotesi alternativa è tipo bidirezionale, si adotterà la seguente regola di decisione:

$$|t| \leq t_{a/2} \qquad \text{si accetta } \mathrm{H_o}$$

$$|t| > \mathrm{t_{a/2}} \qquad \text{si rifiuta } \mathrm{H_o}$$

In base ai dati dell'esercizio, si osserva che le varianze delle popolazioni sono uguali ma incognite e quindi è necessario calcolare la varianza pooled S_p^2 sulla base delle varianze campionarie:

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2} = \frac{50 \cdot 100 + 40 \cdot 90}{88} = 97{,}727$$

Di conseguenza la statistica test sarà pari a :

$$t = \frac{104{,}3 - 97{,}3}{\sqrt{97{,}727(\frac{1}{51} + \frac{1}{41})}} = \frac{7}{\sqrt{4{,}3}} = 3{,}3757$$

Poiché a=0,02 e g=51+41-2=90, si avrà:

$$t_{\frac{a}{2};g} = t_{0{,}01;90} = 2{,}3685$$

Essendo:

$$|t| > \mathrm{t}_{a/2} \quad \rightarrow \quad 3{,}3757 > 2{,}3685 \quad \rightarrow \quad \text{si rifiuta } \mathrm{H}_o$$

Si conclude affermando che, con un livello di confidenza del 98%, si rifiuta l'ipotesi che le medie delle due popolazioni siano uguali.

RIFERIMENTI PER LO STUDIO

MOODLE (http://e-economiaegiurisprudenza.uniparthenope.it/moodle/course/view.php?id=567)-slides:“Lez.6 Campioni indipendenti”

LIBRO DI TESTO (Borra, Di Ciaccio, Statistica, 4ed.): Cap. 14

Eserc.4.6.22 (**DIFFICOLTÀ: AVANZATA)**

Un tour operator nel 2014 ha intervistato un campione di 100 individui ed ha riscontrato che 12 erano suoi clienti. Successivamente nel 2018 ha ricondotto nuovamente l'indagine di mercato su un campione differente composto da 140 individui ed ha riscontrato che il numero di clienti è pari a 21. Si può ipotizzare, con un livello di a=6%, che la QUOTA di mercato del tour operator è aumentata tra il 2014 ed il 2018?

SVOLGIMENTO

L'esercizio richiede di svolgere un test d'ipotesi sulla differenza tra le quote di mercato rilevate negli anni 2018 e 2014. In particolare, considerando π_1 come la quota di mercato nel 2018 (pari a 21/140=0,15) e π_2 come la quota di mercato nel 2014 (pari a 12/100=0,12) si dovrà svolgere un test d'ipotesi sulla differenza tra le proporzioni delle due popolazioni (quote di mercato in anni diversi), il test avrà il seguente sistema d'ipotesi:

$$\mathrm{H_o}: \pi_1 = \pi_2 \qquad \rightarrow \qquad \mathrm{H_o}: \pi_1 - \pi_2 = 0$$

$$\mathrm{H_1}: \pi_1 > \pi_2 \qquad \rightarrow \qquad \mathrm{H_1}: \pi_1 - \pi_2 > 0$$

In particolare

si utilizzerà la statistica test Z:

$$Z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\bar{X}_p(1 - \bar{X}_p)(\frac{1}{n_1} + \frac{1}{n_2})}}$$

Dove $\bar{X}_1$ e $\bar{X}_2$ sono le stime campionarie delle due proporzioni π_1 e π_2 (medie di popolazioni bernoulliane) mentre $\bar{X}_p$ rappresenta lo stimatore congiunto di π :

$$\bar{X}_\mathrm{p} = \frac{n_1\bar{X}_1 + n_2\bar{X}_1}{n_1 + n_2}$$

Considerando che l'ipotesi alternativa è tipo unidirezionale, la regola di decisione sarà:

$$Z \leq z_{a/2} \quad \text{si accetterà } \mathrm{H_o}$$
$$Z > \mathrm{z_{a/2}} \quad \text{si rifiuterà } \mathrm{H_o}$$

In base ai dati dell'esercizio, si calcola la stima della proporzione congiunta pari a:

$$\bar{X}_\mathrm{p} = \frac{140 \cdot 21 + 100 \cdot 12}{140 + 100} = 0{,}1375$$

Di conseguenza, il valore della statistica test sarà :

$$Z = \frac{0{,}15 - 0{,}12}{\sqrt{0{,}1375(1 - 0{,}1375)(\frac{1}{140} + \frac{1}{100})}} = \frac{0{,}03}{\sqrt{0{,}1186 \cdot 0{,}01714}} = \frac{0{,}03}{\sqrt{0.002033036}} = 0{,}4435$$

Poiché a=0,06, si avrà:

$$z_a = z_{0,06} = 1{,}555$$

Essendo:

$$Z < z_a \quad \rightarrow \quad \text{si accetta } \mathrm{H_o}$$

Si conclude affermando che, con un livello di confidenza del 94%, si accetta l'ipotesi che le quote di mercato delle due popolazioni siano uguali (non si è verificata una variazione significativa tra i due periodi considerati).

RIFERIMENTI PER LO STUDIO

MOODLE (http://e-economiaegiurisprudenza.uniparthenope.it/moodle/course/view.php?id=567)-slides:"Lez.6 Campioni indipendenti"

LIBRO DI TESTO (Borra, Di Ciaccio, Statistica, 4ed.): Cap. 14

Eserc.4.7.23 (**DIFFICOLTÀ: MEDIA)**

Si vuole verificare la variabilità del prezzo dei pernottamenti in albergo in due città d'arte. Nella prima città su un campione di n_1=41 alberghi si calcola una varianza campionaria corretta pari a S_1^2=218,5; nella seconda città, su un campione di n_2=31 alberghi, la varianza camp. corretta risulta pari a S_2^2=208,4. Si vuole verificare se la variabilità del prezzo è diversa tra le due città con un 1-α=99%.

__SVOLGIMENTO__

L'esercizio richiede di verificare l'ipotesi di uguaglianza tra le varianze di due popolazioni (ipotesi di omoschedasticità). Nel caso del test tra varianze le due ipotesi si esprimono in termini di rapporto e, considerando la traccia dell'esercizio, il test avrà il seguente sistema d'ipotesi:

$$\mathrm{H_o}: \sigma_1^2 = \sigma_2^2 \qquad \rightarrow \qquad \mathrm{H_o}: \frac{\sigma_1^2}{\sigma_2^2} = 1$$

$$\mathrm{H_1}: \sigma_1^2 \neq \sigma_2^2 \qquad \rightarrow \qquad \mathrm{H_1}: \frac{\sigma_1^2}{\sigma_2^2} \neq 1$$

Per effettuare il test si considera una statistica test F (F di Fisher) data dal rapporto tra le due varianze campionarie:

$$F = \frac{S_1^2}{S_2^2}$$

Considerando che l'ipotesi alternativa è tipo bidirezionale, la regola di decisione sarà:

$$f_{1-\frac{a}{2};\, g1,g2} \leq F \quad \leq f_{\frac{a}{2};\, g1,g2} \qquad \text{si accetta } \mathrm{H_o}$$

$$F \geq f_{\frac{a}{2};\, g1,g2} \qquad \text{si rifiuta } \mathrm{H_o}$$

$$F \leq f_{1-\frac{a}{2};\, g1,g2} \qquad \text{si rifiuta } \mathrm{H_o}$$

In base ai dati dell'esercizio, il valore della statistica test sarà :

$$F = \frac{218{,}5}{208{,}4} = 1{,}048$$

Considerando un livello 1-a=0,99 ed i gradi di libertà g_1=41-1=40 e g_2=31-1=30, in base alle tavole dei valori critici (sulla coda destra) della F di Fischer, si riscontra che :

$$f_{\frac{a}{2};\, g1,g2} = f_{0{,}005;40\,,30} \ = 2{,}52$$

Per calcolare il valore critico sulla coda sinistra della F di Fisher si utilizza la seguente proprietà

$$f_{1-\frac{a}{2};\, g1,g2} = \frac{1}{f_{\frac{a}{2};\, g2,g1}} = \frac{1}{f_{0{,}005;\, 30{,}40}} = \frac{1}{2{,}40} = 0{,}416$$

Essendo:

$$f_{1-\frac{a}{2};\, g1,g2} \leq F \quad \leq f_{\frac{a}{2};\, g1,g2} \quad \rightarrow \quad 0{,}416 \leq \ 1{,}048 \ \leq \ 2{,}52 \quad \rightarrow \quad \text{si accetta } \mathrm{H_o}$$

Si conclude affermando che, con un livello di confidenza del 99%, si accetta l'ipotesi di uguaglianza della variabilità del prezzo dei pernottamenti in albergo nelle due città d’arte .

<u>RIFERIMENTI PER LO STUDIO</u>

MOODLE (http://e-economiaegiurisprudenza.uniparthenope.it/moodle/course/view.php?id=567)-slides:“Lez.6 Campioni indipendenti”

LIBRO DI TESTO (Borra, Di Ciaccio, Statistica, 4ed.): Cap. 14

Eserc.4.8.24 (DIFFICOLTÀ: AVANZATO)

Secondo alcuni annunci recenti, l'affitto medio mensile degli appartamenti con una camera da letto posizionati nel centro città è di 800 euro. Si sospetta che la media sia in realtà più alta. Si seleziona un campione casuale di 25 annunci riferiti ad appartamenti in centro con una camera da letto e si rileva un affitto medio campionario di 850 euro. Si assuma che la deviazione standard della popolazione sia nota e uguale a 120 euro. Verificare l'ipotesi che l'affitto medio di tutti gli appartamenti posizionati in centro con una camera sia superiore a 800 euro mediante approccio del p-value.

SVOLGIMENTO

si dovrà svolgere un test d'ipotesi della media conoscendo l'ammontare della varianza della popolazione, il test avrà il seguente sistema d'ipotesi:

$$\mathrm{H_0}: \mu = 800$$

$$\mathrm{H_1}: \mu > 800$$

si utilizzerà la statistica test Z:

$$Z = \frac{\overline{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$$

In base ai dati dell'esercizio, la statistica test sarà pari a:

$$Z = \frac{850 - 800}{\frac{120}{\sqrt{25}}} = \frac{50}{24} = 2{,}083$$

Si procede al calcolo del p-value cioè il livello di significatività osservato che misura l'evidenza fornita dai dati campionari contro l'ipotesi H_0 (minore è il p-value, più forte è l'evidenza contro H_0).

$$\mathrm{p - value} = 1 - \mathrm{P}(\mathrm{Z} < 2{,}083) = 1 - 0{,}9814 = 0{,}0186$$

Un p-value pari a 0,0186 indica che si rifiuterà l'ipotesi nulla per qualsiasi valore di a>0,0186.

RIFERIMENTI PER LO STUDIO

MOODLE (http://e-economiaegiurisprudenza.uniparthenope.it/moodle/course/view.php?id=567)-slides:"Lez.6 Campioni indipendenti"

LIBRO DI TESTO (Borra, Di Ciaccio, Statistica, 4ed.): Cap. 13

Sezione.5: Regressione lineare semplice

Eserc.5.1.25 (**DIFFICOLTÀ: MEDIA)**

Nella seguente tabella è riportato, per un campione di 8 famiglie, il reddito annuo e la spesa per vacanze sostenuta nel 2019 (valori in migliaia di euro):

Y_i Spesa per vacanze	X_i Reddito annuale
5	41
2	38
5	48
8	44
3	40
2	34
5	35
2	32

Ipotizzando una relazione di tipo lineare tra la spesa per le vacanze ed il reddito:

a. stimare i coefficienti della retta di regressione
b. calcolare l'indice di bontà di adattamento (R^2)

SVOLGIMENTO

Punto a

Ipotizzando che il reddito familiare e la spesa per le vacanze siano legati da una relazione lineare del tipo:

$$Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$$

Sarà possibile calcolare i parametri della retta di regressione mediante il metodo dei minimi quadrati che permette di ricavare una stima puntuale di $\hat{\beta}_0$ e $\hat{\beta}_1$ mediante le seguenti formule:

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

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

Per procedere al calcolo dei parametri sarà quindi necessario calcolare le medie di X e Y:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{312}{8} = 39 \qquad \bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} = \frac{32}{8} = 4$$

successivamente si stimano COD(X,Y) e DEV(X) in modo da poter calcolare $\widehat{\beta}_1$.

Riportando i calcoli in forma tabellare si ottiene:

Y_i Spesa per vacanze	X_i Reddito annuale	$(y_i - \bar{y})$	$(x_i - \bar{x})$	$(y_i - \bar{y})(x_i - \bar{x})$	$(x_i - \bar{x})^2$
5	41	1	2	2	4
2	38	-2	-1	2	1
5	48	1	9	9	81
8	44	4	5	20	25
3	40	-1	1	-1	1
2	34	-2	-5	10	25
5	35	1	-4	-4	16
2	32	-2	-7	14	49
Totale				**52**	**202**

Quindi la COD (X,Y) risulta pari a 52 mentre la DEV(X) è pari a 202, di conseguenza:

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{Cod(X,Y)}{Dev(X)} = \frac{52}{202} = 0{,}2574$$

Avendo stimato $\widehat{\beta}_1$ è possibile procedere alla stima di $\widehat{\beta}_0$che sarà pari a:

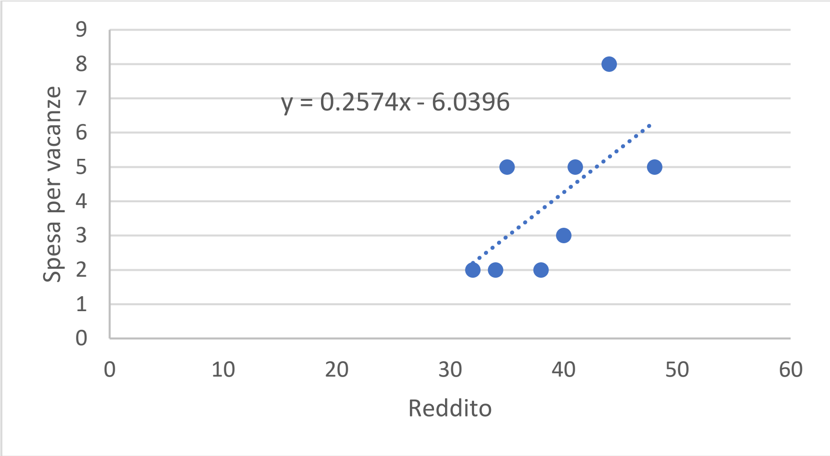
$$\widehat{\beta}_0 = \bar{Y} - \widehat{\beta}_1\bar{X}=4-(0{,}2574\cdot 39) =-6{,}0396$$

Conseguentemente la retta di regressione presenterà la seguente equazione:

Y =- 6.0396+ 0.2574 X

Poiché il coefficiente $\widehat{\beta}_1$ indica la variazione stimata della variabile Y in corrispondenza di una variazione unitaria della variabile X, dati empirici dell’esercizio risulta quindi che per un aumento di reddito di 1000 euro (poiché i dati sono espressi in migliaia) si prevede un aumento della spesa per vacanze pari a257,4 euro.

È anche possibile rappresentare graficamente i dati empirici e la retta stimata:



Punto b

Per calcolare la bontà di adattamento del modello si procederà al calcolo dell'indice R^2 che prevede l'applicazione della seguente formula:

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

Dove SQR indica la Somma dei quadrati della regressione mentre SQT indica la Somma totale dei quadrati.

Sarà quindi necessario calcolare i valori teorici di Y sulla base della retta di regressione stimata in precedenza (punto a) e ricordando che la media di Y è pari a 4:

$$Y = -6.0396 + 0.2574\,X$$

Y_i Spesa per vacanze	X_i Reddito annuale	$\hat{Y}_i$ valori previsti dal modello	$\hat{y}_i - \bar{y}$	$(\hat{y}_i - \bar{y})^2$	$y_i - \bar{y}$	$(y_i - \bar{y})^2$
5	41	4.514	0.514	0.264	1	1
2	38	3.742	-0.258	0.067	-2	4
5	48	6.316	2.316	5.362	1	1
8	44	5.286	1.286	1.654	4	16
3	40	4.256	0.256	0.066	-1	1
2	34	2.712	-1.288	1.659	-2	4
5	35	2.969	-1.031	1.062	1	1
2	32	2.197	-1.803	3.250	-2	4
Totale				13,383		32

Essendo SQR=13,383 e SQT=32, si avrà:

$$R^2 = \frac{\mathrm{SQR}}{\mathrm{SQT}} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = \frac{13{,}383}{32} = 0{,}418$$

Un valore intorno di R^2= 0,40 indica un discreto adattamento del modello ai dati empirici.

È possibile pervenire allo stesso risultato anche utilizzando la formula alternativa che prevede l'utilizzo della Somma dei quadrati degli errori (SQE):

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

Y_i Spesa per vacanze	$\widehat{Y_i}$ valori previsti dal modello	$y_i - \hat{y}_i$	$(y_i - \hat{y}_i)^2$
5	4.514	0.486	0.236
2	3.742	-1.742	3.035
5	6.316	-1.316	1.732
8	5.286	2.714	7.366
3	4.256	-1.256	1.578
2	2.712	-0.712	0.507
5	2.969	2.031	4.125
2	2.197	-0.197	0.039
Totale			**18,617**

Quindi considerando che SQE è uguale a 18,617 mentre SQT è pari a 32, si ottiene:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = 1 - \frac{18{,}617}{32} = 0{,}418$$

<u>RIFERIMENTI PER LO STUDI</u>

MOODLE (http://e-economiaegiurisprudenza.uniparthenope.it/moodle/course/view.php?id=567)-slides:"Lez.7 Regressione Lineare semplice"

LIBRO DI TESTO (Borra, Di Ciaccio, Statistica, 4ed.): Cap. 15

Eserc.5.3.26 (**DIFFICOLTÀ: AVANZATA)**

Da febbraio a settembre del 2019 sono stati osservati i seguenti caratteri statistici: “numero di visitatori agli scavi di Pompei” (in migliaia) e “numero di giornate piovose”

mese	numero visitatori mensili	giornate piovose nel mese
febbraio	14	8
marzo	11	19
aprile	12	15
maggio	18	12
giugno	32	6
luglio	40	3
agosto	22	8
settembre	20	9

Ipotizzando una relazione lineare tra i due caratteri e considerando che lo std. error. del coefficiente β_0 $\mathrm{s(B_0)} = 4.920$, mentre per il coefficiente β_1 risulta $\mathrm{s(B_1)} = 0.444$:

a. costruire un intervallo di confidenza per i coefficienti di regressione (a=1%)
b. testare l’ipotesi che i coefficienti siano significativamente diversi da zero (a=5%)

SVOLGIMENTO

Punto a

Sarà necessario calcolare una stima puntuale dei parametri della retta di regressione mediante il metodo dei minimi quadrati:

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1\bar{X}$$

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

Dopo aver ottenuto una stima puntuale di entrambi i parametri sarà possibile calcolare la stima intervallare. Considerando la distribuzione campionaria degli stimatori dei due parametri:

$$B_0 \sim N\left(\beta_0; \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)\right)$$

$$B_1 \sim N\left(\beta_1; \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)$$

Poiché la quantità σ^2 è incognita, si utilizza il corrispondente stimatore corretto:

$$s^2 = \frac{\sum_{i=1}^{n}(\hat{e}_i - E(\hat{e}_i))^2}{n-2} = \frac{\sum_{i=1}^{n}(\hat{e}_i)^2}{n-2}$$

Conseguentemente per la costruzione dell'intervallo di confidenza si utilizzerà la variabile t di student (con g=n-2), per il coefficiente β_1 si ottiene:

$$P\left(B_1 - t_{\alpha/2;n-2} \cdot s(B_1) < \beta_1 < B_1 + t_{\alpha/2;n-2} \cdot s(B_1)\right) = 1 - \alpha$$

Dove con $\boldsymbol{s(B_1)}$si fa riferimento allo standard error dello stimatore di β_1:

$$s(B_1) = \sqrt{s^2 \Big/ \sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Analogamente per il coefficiente β_0

$$P\left(B_0 - t_{\alpha/2;n-2} \cdot s(B_0) < \beta_0 < B_0 + t_{\alpha/2;n-2} \cdot s(B_0)\right) = 1 - \alpha$$

Dove con $s(B_0)$ si fa riferimento allo standard error dello stimatore di β_0:

$$s(B_0) = \sqrt{s^2\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)}$$

Il primo step consisterà nella stima puntuale dei coefficienti, sarà quindi necessario calcolare le medie di X e Y:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{80}{8} = 10 \qquad \bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} = \frac{169}{8} = 21{,}125$$

successivamente si stimano COD(X,Y) e DEV(X) in modo da poter calcolare $\hat{\beta}_1$, riportando i calcoli in forma tabellare si ottiene:

Yi numero visitatori mensili	Xi giornate piovose nel mese	$(y_i - \bar{y})$	$(x_i - \bar{x})$	$(y_i - \bar{y})(x_i - \bar{x})$	$(x_i - \bar{x})2$
14	8	-7.125	-2	14.25	4
11	19	-10.125	9	-91.125	81
12	15	-9.125	5	-45.625	25
18	12	-3.125	2	-6.25	4
32	6	10.875	-4	-43.5	16
40	3	18.875	-7	-132.125	49
22	8	0.875	-2	-1.75	4
20	9	-1.125	-1	1.125	1
Totale				-305	184

Quindi la COD (X,Y) risulta pari a -305 mentre la DEV(X) è pari a 184, di conseguenza:

$$\hat{\beta}_1 = \frac{Cod(X,Y)}{Dev(X)} = -\frac{305}{184} = -1{,}6576 \qquad \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1\bar{X} = 21{,}125 - (-1{,}6576 \cdot 10) = 37{,}701$$

Considerando che gli std.err sono pari a $\mathrm{s(B_1)} = 0.444$ mentre $\mathrm{s(B_0)} = 4.920$ e che essendo 1-a=0,99 si ottiene:

$$t_{\alpha/2;n-2} = t_{0,005;6} = 3{,}7074 \qquad \text{(g=8-2=6)}$$

L' intervallo di confidenza di β_1 sarà pari a:

$$P\left(B_1 - t_{\alpha/2;n-2} \cdot s(B_1) < \beta_1 < B_1 + t_{\alpha/2;n-2} \cdot s(B_1)\right) = 1 - \alpha$$

$$P(-1{,}6576 - 3{,}7074 \cdot 0.444 < \beta_1 < -1{,}6576 + 3{,}7074 \cdot 0.444) = 0{,}99$$

$$P(-3.304 < \beta_1 < -0.0115) = 0{,}99$$

Mentre quello di β_0:

$$P\left(B_0 - t_{\alpha/2;n-2} \cdot s(B_0) < \beta_0 < B_0 + t_{\alpha/2;n-2} \cdot s(B_0)\right) = 1 - \alpha$$

$$P(37{,}701 - 3{,}7074 \cdot 4.920 < \beta_0 < 37{,}701 + 3{,}7074 \cdot 4.920) = 0{,}99$$

$$P(19.4606 < \beta_0 < 55.9414) = 0{,}99$$

Punto b

Per testare l'ipotesi che i coefficienti siano significativamente diversi da zero si imposterà un test con il seguente sistema d'ipotesi:

$$H_0: \beta_1 = 0 \qquad\qquad H_0: \beta_0 = 0$$

$$H_1: \beta_1 \neq 0 \qquad\qquad H_1: \beta_0 \neq 0$$

Per effettuare il test si utilizzerà una statistica test T:

$$T = \frac{B_1}{s(B_1)} \qquad\qquad T = \frac{B_0}{s(B_0)}$$

Dove $s(B_1)$ e $s(B_0)$ sono i rispettivi std.error dei due coefficienti:

$$s(B_1) = \sqrt{s^2 \Big/ \sum_{i=1}^{n} (x_i - \bar{x})^2} \qquad\qquad s(B_0) = \sqrt{s^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}\right)}$$

Considerando che l'ipotesi alternativa è tipo bidirezionale, si adotterà la seguente regola di decisione:

$$|t| \leq t_{a/2} \qquad \text{si accetta } H_o$$

$$|t| > t_{a/2} \qquad \text{si rifiuta } H_o$$

In base ai dati campionari si calcolerà la seguente statistica test (per il coefficiente β_1):

$$T_{(\beta_1)} = -\frac{1{,}6576}{0{,}444} = -3{,}733$$

Poiché a=0,05, si avrà:

$$t_{\frac{a}{2};g} = t_{0,025;6} = 2{,}4469$$

Essendo:

$$|T| > \mathrm{t}_{\mathrm{a}/2} = 3{,}733 > 2{,}4469 \quad \rightarrow \quad si\ rifiuta\ H_o$$

Con un livello di confidenza del 95%, si rifiuta l'ipotesi che il coefficiente $\beta_1$sia pari a zero e quindi si afferma che sussiste una relazione statisticamente significativa tra giornate piovose e numero visitatori.

Analogamente per quanto riguarda β_0

$$T_{(\beta_0)} = \frac{37{,}701}{4{,}92} = 7{,}663$$

Anche per questo parametro si rifiuterà l'ipotesi nulla in quanto

$$|T| > \mathrm{t}_{\mathrm{a}/2} = 7{,}663 > 2{,}4469 \quad \rightarrow \quad si\ rifiuta\ H_o$$

Quindi, con un livello di confidenza del 95%, si rifiuta l'ipotesi nulla che il coefficiente β_0 sia pari a zero.

<u>RIFERIMENTI PER LO STUDIO</u>

MOODLE (http://e-economiaegiurisprudenza.uniparthenope.it/moodle/course/view.php?id=567)-slides:"Lez.8 Inferenza su parametri regressione"

LIBRO DI TESTO (Borra, Di Ciaccio, Statistica, 4ed.): Cap. 16