Università degli Studi di Bologna - Facoltà di Economia, sede di Forlì

Metodi statistici per l'economia - Prof. A Capitanio

ESERCIZI RISOLTI (seconda parte)

Esercizio 1

Una compagnia di assicurazioni vuole valutare l'entità media delle richieste di risarcimento danni per incidenti automobilistici. Un'indagine svolta su di un campione di 25 richieste ha dato i seguenti risultati (con X si indica la variabile "richiesta di risarcimento in migliaia di euro")

$$\sum_{i=1}^{25} x_i = 112.12 \qquad \sum_{i=1}^{25} x_i^2 = 629.89$$

a) Stimare l'entità media delle richiesta e la varianza delle richieste di risarcimento, giustificando la scelta degli stimatori usati;

Ipotizzando che X abbia distribuzione gaussiana:

b) calcolare l'intervallo di confidenza al 95% per la richiesta media di risarcimento, commentando i passaggi;

a) saggiare, ad un livello di significatività $\alpha = 0.05$, l'ipotesi $H_0 : \mu = 3$ contro l'alternativa $H_1 : \mu > 3$, commentando i passaggi.

Soluzione

a)Possiamo usare gli stimatori media campionaria $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ e varianza corretta $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2$, in quanto sono stimatori corretti dei due parametri da stimare. In corrispondenza del campione osservato ci forniscono le seguenti stime:

$$\bar{x} = \frac{112.12}{25} = 4.485 \qquad s^2 = \frac{Dev(X)}{n-1} = \frac{\sum_{i=1}^{25} x_i^2 - 25\bar{x}^2}{24} = \frac{127.054}{24} = 5.294$$

b) La varianza non è nota e la numerosità campionaria è bassa. Tenendo conto dell'ipotesi di normalità distributiva del carattere X "*richiesta di risarcimento in migliaia di euro*", per il calcolo degli estremi dell'intervallo di confidenza si possono usare le espressioni

$$I_1 = \bar{x} - t_{(n-1),\alpha/2}\frac{s}{\sqrt{n}} \quad \text{e} \quad I_2 = \bar{x} + t_{(n-1),\alpha/2}\frac{s}{\sqrt{n}}$$

dove *s* è lo scarto corretto e $t_{(n-1);\alpha/2}$ va calcolato con riferimento alla distribuzione *t* di Student.

$s^2 = 5.294$, da cui $s = \sqrt{5.294} = 2.301$.

Dalla tavola della *t* di Student si ricava $t_{24,0.025} = 2.0639$ e, poiché

$$\bar{x} - t_{n-1;\frac{\alpha}{2}}\frac{s}{\sqrt{n}} = 4.485 - 0.948 = 3.537$$

$$\bar{x} + t_{n-1;\frac{\alpha}{2}}\frac{s}{\sqrt{n}} = 4.485 + 0.948 = 5.433$$

l'intervallo di confidenza cercato è

[3.537, 5.433]

c) Si tratta di un test di ipotesi unilaterale, e il sistema di ipotesi da considerare è

$$\begin{cases} H_0 : \mu = \mu_0 = 3 \\ H_1 : \mu > \mu_0 = 3 \end{cases}$$

Tenendo conto dell'ipotesi di normalità distributiva per il carattere X e del fatto che la varianza non è nota, la statistica test da usare è

$$t = \frac{\overline{x} - \mu_0}{s/\sqrt{n}}$$

che sotto l'ipotesi H_0 ha distribuzione t_{n-1}. La regione di rifiuto $R = \left\{ t : t_{oss} \geq t_{(n-1),\alpha} \right\}$ è legata alla sola coda destra della distribuzione.

$t_{oss} = \frac{4.485 - 3}{2.301/5} = 3.228$, $t_{24,0.05} \cong 1.7109$ e quindi al prefissato livello di significatività l'ipotesi H_0 è da rifiutare in favore di H_1. Il test è dunque significativo: la richiesta media di risarcimento osservata non è maggiore di 3 per il solo effetto dell'errore di campionamento.

Esercizio 2

Il processo di riempimento delle confezioni di pasta di una azienda non è perfetto: il peso dichiarato è 500 grammi, ma è plausibile che vi siano confezioni con un peso superiore e altre con un peso inferiore. Un'associazione per la tutela del consumatore vuole verificare se il macchinario che riempie le confezioni è veramente tarato su 500 grammi oppure è tarato su un peso inferiore.
Viene selezionato casualmente un campione di 40 confezioni, e ne viene pesato il contenuto; indicato con *X* il "peso in grammi di una confezione". Si osserva

$$\sum_{i=1}^{40} x_i = 19840.02 \qquad \sum_{i=1}^{40} x_i^2 = 9841173.7$$

a)Ipotizzando la normalità distributiva del carattere X, si può affermare, ad un livello di significatività dell'1% che il macchinario è tarato su un peso inferiore?

Soluzione

Si tratta di mettere a confronto le due ipotesi

$$\begin{cases} H_0 : \mu = \mu_0 = 500 \\ H_1 : \mu < \mu_0 = 500 \end{cases}$$

La statistica test da utilizzare è $t = \frac{\overline{x} - \mu_0}{\frac{s}{\sqrt{n}}}$, che andrà poi confrontata con $t_{\alpha;(n-1)}$. Si rifiuta l'ipotesi nulla se il valore osservato per la statistica test è inferiore a $-t_{\alpha;(n-1)}$

$$\overline{x} = \frac{19840.02}{40} = 496 \qquad s^2 = \frac{Dev(X)}{n-1} = \frac{9841173.70 - 40 \times 496^2}{39} = 13.18$$

$$t_{oss} = \frac{\overline{x} - \mu_0}{\frac{s}{\sqrt{n}}} = \frac{496 - 500}{\frac{3.63}{\sqrt{40}}} = -6.98$$

Dalle tavole si ottiene $t_{0.01;39} = 2.426$, e quindi il test è significativo: si rifiuta l'ipotesi nulla. Al prefissato livello di significatività è possibile affermare che il macchinario è tarato su un peso inferiore a 500 gr.

Esercizio 3

Durante l'ultimo anno un'azienda ha introdotto l'orario "flessibile" (ogni impiegato può, entro certi limiti, scegliere l'orario di lavoro più adatto alle sue esigenze). Il numero medio di giorni di assenza per impiegato, nei tre anni precedenti, è stato di 6.3 giorni all'anno. Per verificare se l'introduzione dell'orario flessibile ha ridotto l'assenteismo, come alcuni dirigenti hanno sostenuto, viene estratto un campione casuale di 100 impiegati, e viene registrato il numero di giorni di assenza di ciascuno nel corso dell'ultimo anno. Indicato con X il numero di giorni di assenza per ciascun impiegato si e osservato

$$\sum_{i=1}^{100} x_i = 550 \qquad \sum_{i=1}^{100} x_i^2 = 3866$$

a) Stimare, giustificando la scelta dello stimatore, la varianza dei giorni di assenza.
b) Si può affermare, ad un livello di significatività pari a 0.5%, che l'orario flessibile riduce l'assenteismo? Rispondere commentando sia il risultato sia la procedura usata.

Soluzione

a)Possiamo usare la varianza corretta $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2$, in quanto è stimatore non distorto della varianza di popolazione. In corrispondenza del campione osservato si osserva la stima

$$s^2 = \frac{Dev(X)}{n-1} = \frac{\sum_{i=1}^{100} x_i^2 - 100\bar{x}^2}{99} = \frac{3866 - 3025}{99} = \frac{841}{99} = 8.49 \quad \text{in quanto } \bar{x} = \frac{550}{100} = 5.5$$

b) Il sistema di ipotesi da considerare è

$$\begin{cases} H_0 : \mu = 6.3 \\ H_1 : \mu < 6.3 \end{cases}$$

e quindi si tratta di un test ad una coda. Tenendo conto della non modesta numerosità campionaria, possiamo usare l'approssimazione alla normale, e usare la statistica test

$$Z = \frac{\bar{X} - \mu_0}{s/\sqrt{n}}$$

che sotto l'ipotesi H_0 ha distribuzione $N(0;1)$. La regione di rifiuto $R = \left\{z : z_{oss} < -z_\alpha\right\}$ è legata alla sola coda sinistra della distribuzione.

$z_{oss} = \frac{5.5 - 6.3}{2.91/10} = -2.74$, $\quad z_{0.005} \cong 2.575$ e quindi il valore osservato della statistica test cade nella zona di rifiuto. Al prefissato livello di significatività l'ipotesi H_0 è da rifiutare in favore di H_1. Il test è dunque significativo: Il numero medio osservato di giorni di assenza dal lavoro non è inferiore a 6.3 per il solo effetto dell'errore di campionamento, e quindi si può supporre che l'orario flessibile abbia effetti positivi sull'assenteismo.

Esercizio 4

I dati che seguono si riferiscono, con riferimento al 2004, ad un campione casuale semplice di 150 creditori di una banca che sono risultati insolventi, non avendo restituito il prestito ad essi accordato; con X viene indicato il credito non recuperato espresso in migliaia di euro.

$$\sum_{i=1}^{n} x_i = 6232.40 \qquad \sum_{i=1}^{n} x_i^2 = 402709.08$$

a) Stimare il credito medio non recuperato, e la varianza del credito non recuperato, giustificando la scelta degli stimatori usati.
b) Nel biennio 2002-2003 il credito mediamente non recuperato è stato di 45.25 migliaia di euro. Nel 2004 è stato modificato il sistema di valutazione del rischio di credito. Si può affermare, ad un livello di significatività del 5%, che il sistema di valutazione introdotto ha ridotto l'ammontare medio del credito non recuperato? Giustificare il procedimento di calcolo e commentare il risultato.

Soluzione

a) Le due quantità possono essere stimate rispettivamente con lo stimatore media aritmetica campionaria e varianza campionaria corretta. Si tratta di stimatori non distorti e consistenti.

$$\overline{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{6232.40}{150} = 41.55$$

$$s_n^2 = \frac{Dev(X)}{n-1} = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n\overline{x}^2\right) = \frac{402709.08 - 150 \times 41.55^2}{149} = \frac{143757.01}{149} = 964.81$$

b) Il sistema di ipotesi da considerare è:

$$\begin{cases} H_0 : \mu = 45.25 \\ H_1 : \mu < 45.25 \end{cases}$$

e quindi si tratta di un test ad una coda. Non viene fatta alcuna ipotesi riguardo la distribuzione del carattere: tenendo conto della elevata numerosità campionaria, in virtù del teorema limite centrale possiamo usare l'approssimazione alla normale, e usare la statistica test

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

che sotto l'ipotesi H_0 ha distribuzione $N(0;1)$. La regione di rifiuto $R = \{z : z_{oss} < -z_\alpha\}$ è legata alla sola coda sinistra della distribuzione.

$z_{oss} = \dfrac{41.55 - 45.25}{31.06/\sqrt{150}} = -1.4591$, $\quad z_{0.05} \cong 1.645$ e quindi il valore osservato della statistica test non cade nella zona di rifiuto. Al prefissato livello di significatività l'ipotesi H_0 non può essere rifiutata in favore di H_1, e quindi il test non è significativo. Il nuovo sistema di valutazione introdotto non ha ridotto l'ammontare medio del credito non recuperato, che risulta diverso da quello del biennio 2002-2003 per il solo effetto dell'errore di campionamento.

Esercizio 5

Viene effettuato un sondaggio per prevedere quale fra due candidati alla carica di sindaco di una città vincerà il ballottaggio. Indichiamo con A e B i due candidati. Vengono fatte 200 interviste,

nelle quali all'intervistato viene chiesto di esprimere la propria preferenza; il candidato B riceve 105 preferenze.

a) Scegliere il modello distributivo più opportuno per la variabile osservata, e descriverlo.
b) Stimare la probabilità che il candidato B diventi sindaco, giustificando la scelta dello stimatore usato.
c) Calcolare l'intervallo di confidenza al 90% per la probabilità che il candidato B diventi sindaco, giustificando la procedura di calcolo scelta.
d) Possiamo affermare con certezza che il candidato B vincerà le elezioni? Motivare la risposta.
e) Anche durante le elezioni precedenti i due candidati erano arrivati al ballottaggio, e aveva vinto il candidato B con il 55% delle preferenze. Si può affermare ad un livello di significatività del 5% che il candidato B ha perso popolarità?

Soluzione

a) Stiamo osservando il valore assunto dalla variabile *X*="Preferenza per uno dei due candidati", che può assumere solo le due modalità "Preferisco il candidato A" e "Preferisco il candidato B". Il modello distributivo che si usa in contesti di questo tipo è la variabile aleatoria di Bernoulli: il carattere X che osserviamo può assumere solo due modalità: "*Preferenza per il candidato B",* codificata da X=1 e "*Preferenza il candidato A"* , codificata da X=0.
b) La probabilità che il candidato B diventi sindaco può essere stimata attraverso lo stimatore frequenza campionaria, che è uno stimatore non distorto e consistente.
Sulla base del campione otteniamo la stima
$f_n = \frac{105}{200} = 0.525$
c) La distribuzione del carattere *X* è *Ber*(*p*). L'ampiezza campionaria è sufficientemente ampia per poter invocare il Teorema centrale del limite e calcolare per *p* un intervallo di confidenza asintotico, la cui espressione generale è:

$$\left[f_n - z_{\frac{\alpha}{2}} \sqrt{\frac{f_n(1-f_n)}{n}}; f_n + z_{\frac{\alpha}{2}} \sqrt{\frac{f_n(1-f_n)}{n}} \right]$$

$1-\alpha = 0.90 \quad \alpha = 0.10 \quad \frac{\alpha}{2} = 0.05 \quad z_{0.05} = 1.64485 \quad f_{200} = 0.525$

$$\sqrt{\frac{f_n(1-f_n)}{n}} = \sqrt{\frac{0.525(0.475)}{200}} = 0.03531$$

$$f_n - z_{\frac{\alpha}{2}} \sqrt{\frac{f_n(1-f_n)}{n}} = 0.525 - 1.64485 \times 0.03531 = 0.4669$$

$$f_n + z_{\frac{\alpha}{2}} \sqrt{\frac{f_n(1-f_n)}{n}} = 0.525 + 1.64485 \times 0.03531 = 0.5831$$

Dunque l'intervallo di confidenza risulta essere [0.4669; 0.5831].

d) No, non possiamo affermarlo con certezza in quanto stiamo lavorando con dati campionari, e quindi la stima ottenuta al punto b), pari al 52.5% dei voti, è soggetta all'errore di campionamento. Inoltre non possiamo affermarlo neanche con un livello di confidenza del 90%, in quanto il risultato ottenuto al punto c) consente di affermare che, con un livello di confidenza del 90%, il candidato otterrà una percentuale di preferenze che va dal 46.49% (inferiore al 50%) al 58.31%.

e) Si tratta di un test di ipotesi su una proporzione con livello di significatività $\alpha = 0.05$. Le ipotesi messe a confronto sono:

$$\begin{cases} H_0 : p = 0.55 \\ H_1 : p < 0.55 \end{cases}$$

Vista la non modesta numerosità campionaria possiamo considerare la statistica test

$Z = \dfrac{f_n - p_0}{\sqrt{\dfrac{p_0(1-p_0)}{n}}}$, che sotto l'ipotesi H_0 al crescere della numerosità campionaria tende a distribuirsi come una Normale standardizzata. Si rifiuterà H_0 se il valore osservato della statistica test è inferiore a $-z_\alpha = -z_{0.05} = -1.64485$. Utilizzando i risultati ottenuti ai punti precedenti si ha:

$$z_{oss} = \frac{f_n - p_0}{\sqrt{\dfrac{p_0(1-p_0)}{n}}} = \frac{0.525 - 0.55}{0.035178} = -0.71067$$

Il test non è significativo: al livello di significatività del 5% non si può affermare che il candidato B ha perso popolarità rispetto alle elezioni precedenti.

Esercizio 6

La ditta A, che produce transistor, vuole valutare la qualità della propria produzione. A tal fine seleziona un campione casuale di 150 transistor: fra questi 20 risultano difettosi.

a) Stimare la probabilità che un transistor sia difettoso
b) Costruire l'intervallo di confidenza al 90% per tale probabilità

Soluzione

a)Il carattere X che osserviamo può assumere solo due modalità: "*difettoso"* $\rightarrow$ X=1 e "*non difettoso"* $\rightarrow$ X=0 con probabilità rispettivamente *q=(1-p)* e *p*. Dobbiamo stimare *p*; uno stimatore non distorto per tale parametro è la frequenza dei pezzi difettosi, ovvero $f_n = \hat{p} = \dfrac{\text{numero di pezzi difettosi}}{n}$. In corrispondenza del campione osservato si ottiene la stima

$$f_n = \frac{20}{150} = 0.134$$

c) La distribuzione del carattere X è Ber(p). L'ampiezza campionaria è sufficientemente ampia per poter invocare il Teorema centrale del limite e usare l'espressione nota per gli intervalli asintotici.

$$\left[f_n - z_{\frac{\alpha}{2}} \sqrt{\frac{f_n(1-f_n)}{n}}; f_n + z_{\frac{\alpha}{2}} \sqrt{\frac{f_n(1-f_n)}{n}} \right]$$

$1-\alpha = 0.90 \quad \alpha = 0.10 \quad \dfrac{\alpha}{2} = 0.05 \quad z_{0.05} = 1.645$

$$\sqrt{\frac{f_n(1-f_n)}{n}} = \sqrt{\frac{0.134(0.866)}{150}} = 0.028$$

$$f_n - z_{\frac{\alpha}{2}}\sqrt{\frac{f_n(1-f_n)}{n}} = 0.134 - 1.645 \times 0.028 = 0.088$$

$$f_n + z_{\frac{\alpha}{2}}\sqrt{\frac{f_n(1-f_n)}{n}} = 0.134 + 1.645 \times 0.028 = 0.180$$

Dunque l'intervallo di confidenza risulta essere [0.088; 0.180].

Esercizio 7

Un test a scelta multipla è formato da 10 domande: per ciascuna domanda vengono fornite tre risposte (fra le quali solo una è esatta). Calcolare la probabilità che, rispondendo a caso, almeno 8 risposte risultino corrette.

Soluzione

Il numero di risposte corrette rispondendo a caso può essere modellato attraverso una variabile aleatoria binomiale di parametri n=10 e p=1/3, che indichiamo con X. Infatti il problema è riconducibile al calcolo della probabilità di avere almeno 8 "successi" in 10 prove indipendenti, ciascuna con identica probabilità di successo, pari a 1/3.
Si tratta quindi di calcolare $P(X \geq 8) = P(X = 8) + P(X = 9) + P(X = 10)$.

Ricordando che se $X \sim B(n, p)$, allora $P(X = x) = \binom{n}{x} p^x (1-p)^{n-x} = \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x}$, si ha:

$$P(X \geq 8) = \frac{10!}{8!2!}\left(\frac{1}{3}\right)^8\left(\frac{2}{3}\right)^2 + \frac{10!}{9!1!}\left(\frac{1}{3}\right)^9\left(\frac{2}{3}\right)^1 + \frac{10!}{10!0!}\left(\frac{1}{3}\right)^{10}\left(\frac{2}{3}\right)^0 = 45\frac{4}{3^{10}} + 10\frac{2}{3^{10}} + \frac{1}{3^{10}} =$$

$$= \frac{201}{3^{10}} \cong 0.0034$$

Esercizio 8

Il tempo (in mesi) che intercorre fra la laurea e il primo impiego per i laureati in economia è una variabile aleatoria N(8, 6).
a) Calcolare la probabilità che passi più di un anno prima che un neolaureato trovi il primo impiego.

Soluzione

Indichiamo con Y la variabile aleatoria "*Tempo (in mesi) che intercorre fra la laurea e il primo impiego per i laureati in economia*". Sappiamo che Y ha distribuzione Normale di parametri μ=8 e σ^2 = 6.
Dobbiamo determinare $P(Y>12)$, il che equivale a calcolare $1 - P(Y \leq 12) = 1 - F_Y(12) = 1 - \Phi\left(\frac{12-8}{\sqrt{6}}\right) = 1 - \Phi(1.63)$, dove $\Phi(\cdot)$ è la Funzione di Ripartizione di una Normale standardizzata. Dalle tavole ricaviamo $\Phi(1.63) = 0.9484$, e quindi

$$P(Y > 12) = 1 - \Phi(1.63) = 1 - 0.9484 = 0.0516$$

Esercizio 9

Una banca ha finanziato 10 imprese. Sapendo che la probabilità che un'impresa sia insolvente è pari a 0.2, e assumendo che le imprese si comportino in maniera indipendente,

a) Determinare la probabilità che due fra le imprese finanziate siano insolventi.
b) Determinare la probabilità che almeno 2 fra le imprese finanziate siano insolventi.
c) Determinare la probabilità che meno di 9 fra le imprese finanziate siano insolventi.

Soluzione

Nelle ipotesi fatte, la variabile aleatoria X = "*numero di imprese insolventi*" ha distribuzione $B(10;0.2)$. La probabilità cercata è quindi

a) $P(X=2)=\binom{10}{2}0.2^2\times 0.8^8=\frac{10!}{2!8!}0.2^2\times 0.8^8=45\times 0.04\times 0.168=0.302$

b) $P(X\geq 2)=1-P(X<2)=1-P(X=0)-P(X=1)=1-\frac{10!}{0!10!}0.2^0\times 0.8^{10}-\frac{10!}{1!9!}0.2^1\times 0.8^9=$

$=1-0.8^{10}-10\times 0.2\times 0.8^9=0.62419$

c) $P(X<9)=1-P(X\geq 9)=1-P(X=9)-P(X=10)=1-\frac{10!}{9!1!}\times 0.2^9\times 0.8^1-\frac{10!}{10!0!}\times 0.2^{10}\times 0.8^0=$

$=1-10\times 0.2^9\times 0.8-0.2^{10}=0.9999$

Esercizio 10

L'incasso giornaliero (in migliaia di euro) del mio negozio è una variabile aleatoria N(μ=3,σ^2=0.1).

a) Qual è la probabilità che io domani incassi meno di 3500 euro?
b) Calcolare la probabilità che l'incasso sia compreso fra 2900 e 3500 euro.
c) Calcolare la probabilità che l'incasso totale in 30 giorni di attività non superi 95000 euro.
d) Qual è la distribuzione dell'incasso medio giornaliero di trenta giorni di attività?

Soluzione

Indichiamo con Y la variabile aleatoria "incasso giornaliero del mio negozio". Sappiamo che Y ha distribuzione Normale di parametri μ=3 e σ^2 = 0.1.

a) Dobbiamo determinare P(Y≤3.5), il che equivale a calcolare

$P\left(\frac{Y-3}{\sqrt{0.1}}\leq\frac{3.5-3}{\sqrt{0.1}}\right)=P\left(Z\leq 1.582\right)=\Phi(1.582)$, dove Z è una Normale standardizzata. Dalle tavole vediamo che la probabilità cercata è pari circa a 0.9429.

b) Si tratta di calcolare P(2.9<Y≤3.5).

$P(2.9<Y\leq 3.5)=P(Y\leq 3.5)-P(Y\leq 2.9)=F_Y(3.5)-F_Y(2.9)=$

$=\Phi\left(\frac{3.5-3}{\sqrt{0.1}}\right)-\Phi\left(\frac{2.9-3}{\sqrt{0.1}}\right)=\Phi\left(1.582\right)-\Phi\left(-0.316\right)=(1-0.6255)-0.9429=0.5684$

in quanto $\Phi\left(-0.316\right)=1-\Phi\left(0.316\right)=1-0.6255$.

c) L'incasso di ciascun giorno è descritto dal modello $Y_i\sim N(3;0.1)$, e quindi l'incasso totale in 30 giorni corrisponde alla somma $Y_{TOT}=\sum_{i=1}^{30}Y_i$ di 30 variabili aleatorie normali, distribuite come Y,

e che possiamo assumere indipendenti. Per una delle proprietà della distribuzione Normale sappiamo che tale somma ha anch'essa distribuzione Normale, con valore atteso pari alla somma dei valori attesi e varianza pari alla somma delle varianze delle variabili aleatorie addende. Dunque si ha $Y_{TOT} \sim N(90;3)$ e si tratta di calcolare $P(Y_{TOT} \leq 95)$.

$$P(Y_{TOT} \leq 95) = \Phi\left(\frac{95-90}{\sqrt{3}}\right) = \Phi(2.89) = 0.998$$

d) La media aritmetica di *n* variabili casuali Normali indipendenti e con identici parametri è distribuita come una Normale con valore atteso uguale al valore atteso delle variabili addende e varianza pari alla varianza delle variabili addende divisa per *n*. Quindi nel nostro caso si ha

$$Y_i \sim N(3;0.1) \qquad Y_{TOT} = \sum_{i=1}^{30} Y_i \qquad \bar{Y} = \frac{\sum_{i=1}^{30} Y_i}{30} = \frac{Y_{TOT}}{30}$$

$$\bar{Y} \sim N\left(3;\frac{0.1}{30}\right)$$

Esercizio 11

Il numero di pratiche evase giornalmente da un impiegato di un ufficio pubblico può essere descritto dalla variabile aleatoria X così descritta:

X	10	11	12	13	14	15
P(X=x)	0.01	0.15	0.2	0.35	0.24	0.05

a) Calcolare la probabilità che in un giorno l'impiegato completi più di 12 pratiche.
b) Calcolare il numero atteso di pratiche completate in un giorno dall'impiegato.

Soluzione

a) $P(X > 12) = 0.35 + 0.24 + 0.05 = 0.64$

b) $E(X) = \sum x \cdot P(X = x) = 0.01 \times 10 + 0.15 \times 11 + \ldots + 0.05 \times 15 = 12.81$

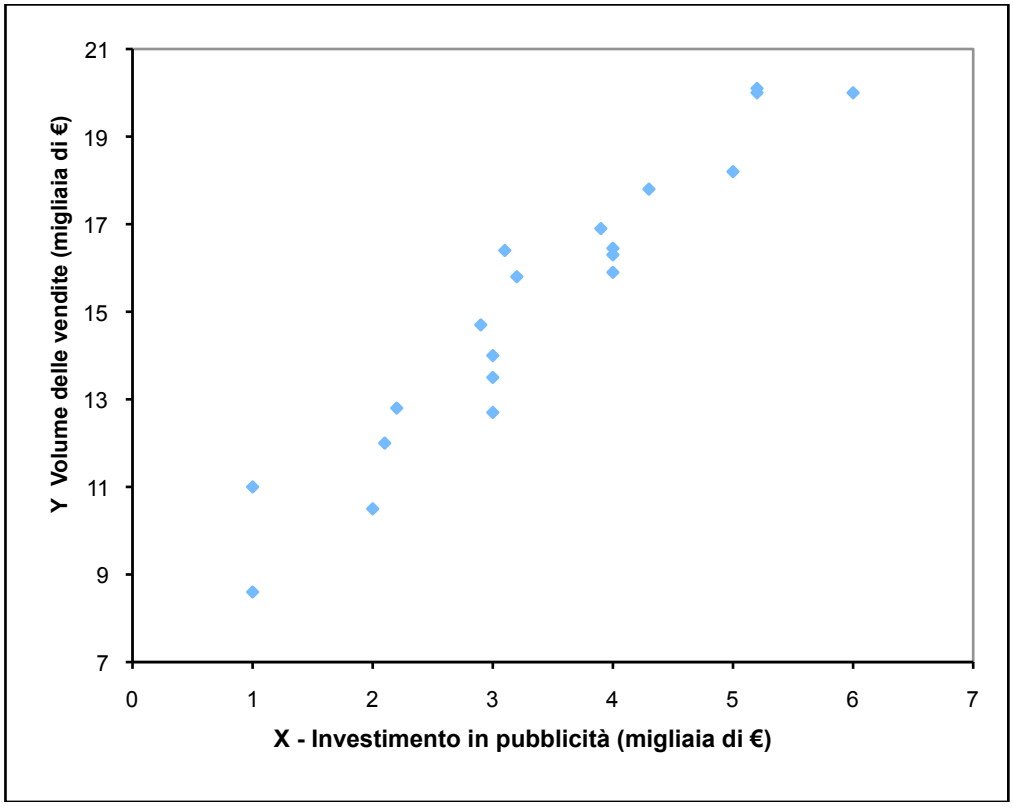
Esercizio 12

Una grande società che possiede una catena di negozi situati in diverse città italiane vuole valutare l'impatto avuto da un investimento in pubblicità su quotidiani locali e attraverso volantini fatto nel mese di settembre. A tal fine seleziona un campione casuale di 20 negozi, situati in città simili per dimensione demografica e per altre caratteristiche che possono influenzare il comportamento d'acquisto. Per ciascun negozio vengono rilevati: investimento in pubblicità (X) e volume delle vendite (Y) fra il 20 settembre e il 30 novembre. Di seguito sono riportati i dati e una loro rappresentazione grafica mediante diagramma di dispersione.

Città (codice)	**X** (Inv. in pubblicità)	**Y** (Volume vendite)
1	1.00	11.00
2	1.00	8.60
3	2.00	10.50
4	2.10	12.00
5	2.20	12.80
6	2.90	14.70
7	3.00	13.50
8	3.00	14.00
9	3.00	12.70
10	3.10	16.40
11	3.20	15.80
12	3.90	16.90
13	4.00	16.30
14	4.00	15.90
15	4.00	16.45
16	4.30	17.80
17	5.00	18.20
18	5.20	20.00
19	5.20	20.10
20	6.00	20.00
Totale	**68.10**	**303.65**
Media	**3.41**	**15.18**

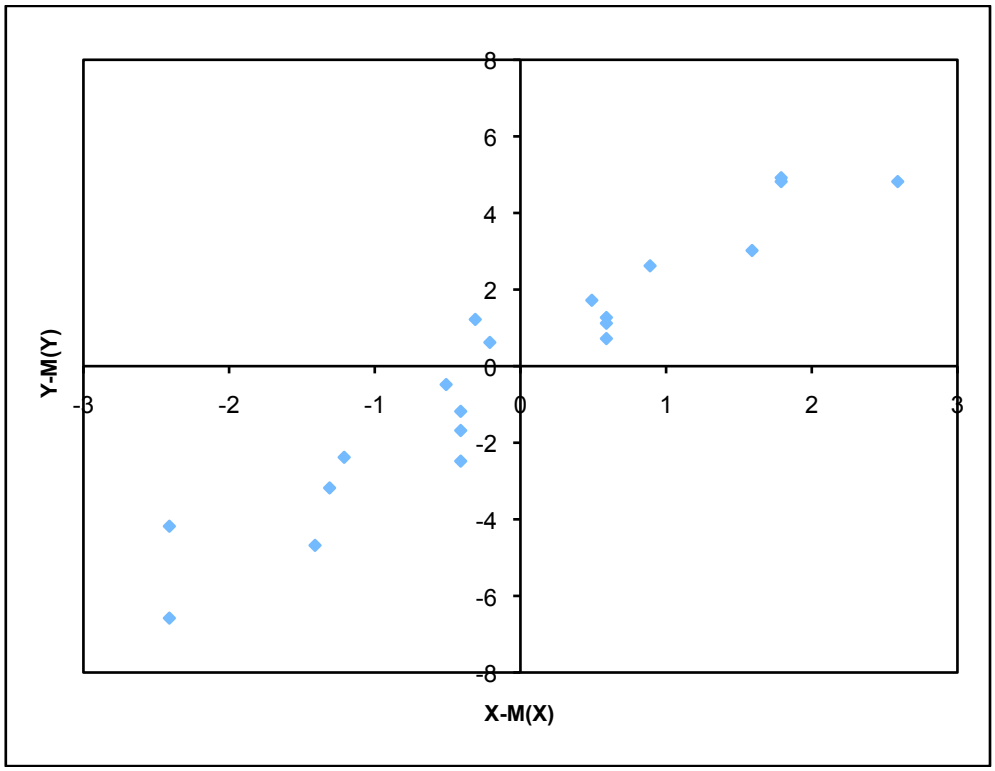
Alcune sintesi dei dati:

$\sum_{i=1}^{n} x_i y_i = 1115.16$ $\sum_{i=1}^{n} x_i^2 = 267.29$ $\sum_{i=1}^{n} y_i^2 = 4814.08$ $\bar{X} = 3.41$ $\bar{Y} = 15.18$



a) Valutare graficamente la connessione fra X e Y.

E' opportuno considerare le variabili "scarto dalla media" $X-\bar{x}$ e $Y-\bar{Y}$



Fra i dati prevale concordanza: a valori più alti della media per una variabile si accompagnano prevalentemente valori più alti della media per l'altra, e a valori più bassi della media per una variabile si accompagnano prevalentemente valori più bassi della media per l'altra.
b) Quantificare l'interdipendenza lineare attraverso il coefficiente di correlazione lineare.

$$\sum_{i=1}^{n} x_i y_i = 1115.16 \qquad \sum_{i=1}^{n} x_i^2 = 267.29 \qquad \sum_{i=1}^{n} y_i^2 = 4814.08$$

$$\bar{X} = 3.41 \qquad \bar{Y} = 15.18$$

$$Dev(X) = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2 = 267.29 - 20 \times 3.41^2 = 35.41$$

$$Dev(Y) = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2 = 4814.08 - 20 \times 15.18^2 = 203.92$$

$$Cod(X,Y) = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i y_i - n\bar{x}\,\bar{y} = 1115.16 - 20 \times 3.41 \times 15.18 = 81.23$$

$$\rho_{XY} = \frac{Cod(X,Y)}{\sqrt{Dev(X)Dev(Y)}} = \frac{81.23}{\sqrt{35.41 \times 203.92}} = 0.96$$

Si riscontra un'elevata correlazione lineare positiva (il valore del coefficiente di correlazione è positivo e molto vicino a 1).

c) Sulla base dei risultati ottenuti, formulare una valutazione riguardo il volume delle vendite che ci si aspetta di ottenere in corrispondenza di un investimento in pubblicità pari a 4500€, e dell'investimento in pubblicità che sarebbe stato necessario per ottenere un volume delle vendite pari a circa 10000€.

Visto che il coefficiente di correlazione è positivo, e che il suo valore assoluto segnala intensa concordanza, in corrispondenza di un investimento in pubblicità pari a 4500€, valore che è più alto della media, ci si aspetta di osservare un volume delle vendite più alto della media, ovvero maggiore di 15180€. Sulla base di analoghe considerazioni si desume che per ottenere un volume delle vendite pari a circa 10000€, che è un valore più basso della media, sarebbe stato necessario fare un investimento in pubblicità inferiore alla media, cioè minore di 3410€.

Sulla base del campione osservato si vogliono stimare i parametri del modello
$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$
che approssima il valore mediamente assunto dal volume delle vendite attraverso una funzione dei valori assunti dall'investimento in pubblicità.

d) Stimare i parametri β_0 e β_1 con gli stimatori dei Minimi Quadrati.
Le espressioni da usare sono:

$$\hat{\beta}_1 = \frac{Cod(X,Y)}{Dev(X)} = \frac{\sum_{i=1}^{n} x_i y_i - n\overline{xy}}{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2} \qquad \text{e} \qquad \hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}$$

e le stime, utilizzando i calcoli fatti al punto b), risultano essere

$$\hat{\beta}_1 = \frac{Cod(X,Y)}{Dev(X)} = \frac{81.23}{35.41} = 2.29 \qquad \text{e} \qquad \hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x} = 15.18 - 2.29 \times 3.41 = 7.37$$

Il coefficiente di regressione è positivo; all'aumentare dell'investimento in pubblicità il volume delle vendite aumenta. In particolare, ad un aumento unitario (1000€) dell'investimento in pubblicità il volume delle vendite aumenta di 2290€.

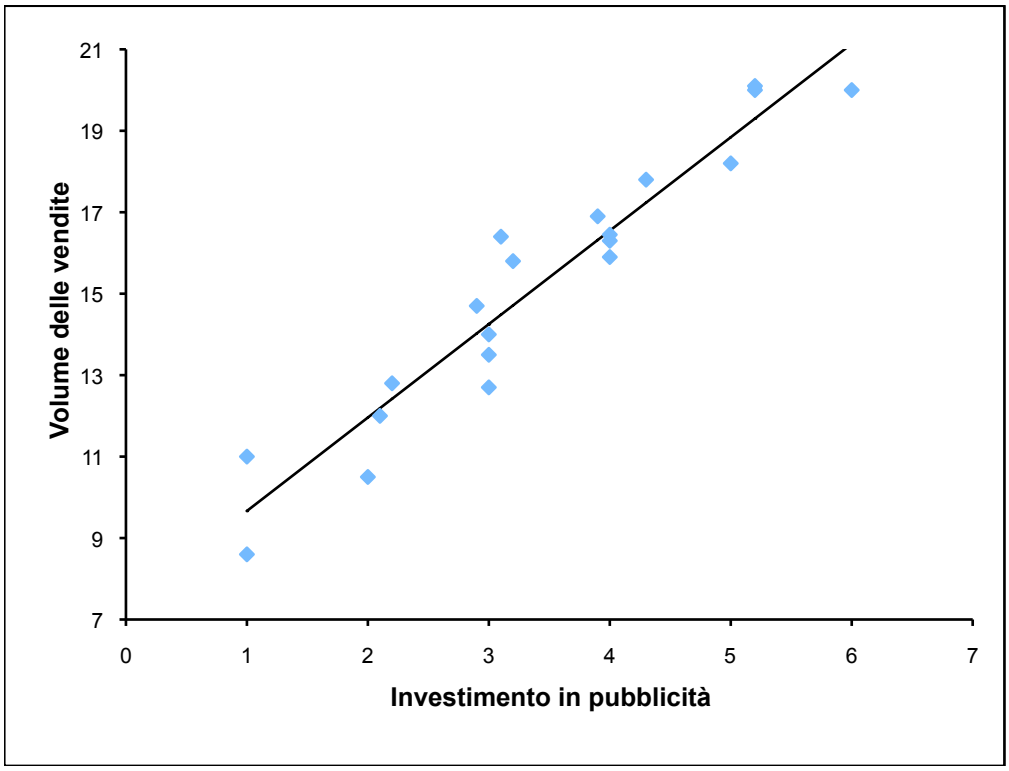
e) Calcolare i valori stimati del volume medio delle vendite in corrispondenza dei valori osservati per l'investimento in pubblicità.

Si usa il modello stimato: $\hat{y}_i = 7.37 + 2.29 x_i$

$\hat{y}_1 = \hat{y}_2 = 7.37 + 2.29 \times 1 = 9.67$
$\hat{y}_3 = 7.37 + 2.29 \times 2 = 11.96$
…
I valori sono riportati nella tabella seguente, che include anche i residui stimati

Città (codice)	**X** (Inv. in pubblicità)	**Y** (Volume vendite)	$\hat{y}$ **valori stimati**	$\hat{e}$ **residui**
1	1.00	11.00	9.67	1.33
2	1.00	8.60	9.67	-1.07
3	2.00	10.50	11.96	-1.46
4	2.10	12.00	12.19	-0.19
5	2.20	12.80	12.42	0.38
6	2.90	14.70	14.02	0.68
7	3.00	13.50	14.25	-0.75
8	3.00	14.00	14.25	-0.25
9	3.00	12.70	14.25	-1.55
10	3.10	16.40	14.48	1.92
11	3.20	15.80	14.71	1.09
12	3.90	16.90	16.32	0.58
13	4.00	16.30	16.55	-0.25
14	4.00	15.90	16.55	-0.65
15	4.00	16.45	16.55	-0.10
16	4.30	17.80	17.24	0.56
17	5.00	18.20	18.84	-0.64
18	5.20	20.00	19.30	0.70
19	5.20	20.10	19.30	0.80
20	6.00	20.00	21.14	-1.14
Totale	**68.10**	**303.65**		**0.00**

f) Rappresentare graficamente la retta stimata



g) Calcolare i residui, e verificare che la loro somma (e quindi la media aritmetica) è uguale a zero.

I residui si stimano con l'espressione $\hat{e}_i = y_i - \hat{y}_i$. I valori sono riportati nella tabella, ed è immediato verificare che la loro somma è zero.

h) Valutare l'intensità della dipendenza lineare di Y da X.

Si usa l'indice di determinazione lineare $R^2 = \frac{Dev_{regr}}{Dev(Y)}$

Per calcolare la devianza di regressione usiamo l'espressione

$Dev_{regr}(Y) = \hat{\beta}_1^2 Dev(X) = 2.29^2 \times 35.41 = 186.35$, da cui

$R^2 = \frac{Dev_{regr}}{Dev(Y)} = \frac{186.35}{203.92} = 0.91$

L'intensità del legame è molto elevata. Il 91% della variabilità del volume delle vendite è spiegato dalla dipendenza lineare dall'investimento in pubblicità.

i) Stimare la varianza σ^2 degli errori ε_i

Uno stimatore non distorto di σ^2 è $s^2 = \frac{Dev_{disp}(Y)}{n-2} = \frac{SQE}{n-2}$

$Dev_{disp}(Y) = Dev(Y) - Dev_{regr}(Y) = 203.92 - 186.35 = 17.57$

(Nota: visto che abbiamo a disposizione i residui, in alternativa si poteva usare l'espressione $Dev_{disp}(Y) = \sum_{i=1}^{n} \hat{e}_i^2$)

Otteniamo $s^2 = \frac{17.57}{18} = 0.9759$

l) Ipotizzando che gli errori ε_i abbiano distribuzione $N(0, \sigma^2)$, calcolare gli intervalli di confidenza al 95% per β_1 e β_0.

Nell'ipotesi fatta è possibile usare le seguenti espressioni:

$$\left[\hat{\beta}_1 - t_{\frac{\alpha}{2}} s(B_1); \hat{\beta}_1 + t_{\frac{\alpha}{2}} s(B_1)\right] \quad \text{e} \quad \left[\hat{\beta}_0 - t_{\frac{\alpha}{2}} s(B_0); \hat{\beta}_0 + t_{\frac{\alpha}{2}} s(B_0)\right]$$

La distribuzione di riferimento è una $t_{n-2} = t_{18}$

$\alpha = 0.05 \quad \alpha/2 = 0.025 \quad t_{\frac{\alpha}{2}} = 2.1009$

$$s(B_1) = \sqrt{\frac{Dev_{disp}}{(n-2)Dev(X)}} = \sqrt{\frac{s^2}{Dev(X)}} = \sqrt{\frac{0.9759}{35.41}} = 0.166$$

$\hat{\beta}_1 - t_{\frac{\alpha}{2}} s(B_1) = 2.29 - 2.1009 \times 0.166 = 1.9453$

$\hat{\beta}_1 + t_{\frac{\alpha}{2}} s(B_1) = 2.29 + 2.1009 \times 0.166 = 2.6428$

Con un livello di confidenza dl 95% il valore del parametro β_1 è compreso nell'intervallo [1.453; 2.6428]

$$s(B_0) = \sqrt{s^2\left[\frac{1}{n} + \frac{\bar{x}^2}{Dev(X)}\right]} = \sqrt{0.9759\left[\frac{1}{20} + \frac{3.41^2}{35.41}\right]} = 0.6069$$

$$\hat{\beta}_0 - t_{\frac{\alpha}{2}} s(B_0) = 7.37 - 2.1009 \times 0.6069 = 6.0962$$

$$\hat{\beta}_0 + t_{\frac{\alpha}{2}} s(B_0) = 7.37 + 2.1009 \times 0.6069 = 8.6462$$

Con un livello di confidenza del 95% il valore del parametro β_0 è compreso nell'intervallo [6.0962; 8.6462]

m) Ipotizzando che gli errori ε_i abbiano distribuzione $N(0,\sigma^2)$, verificare se i coefficienti del modello sono significativamente diversi da zero. (Si consideri un livello di significatività dell'1%)

Consideriamo prima il coefficiente di regressione β_1. Il sistema di ipotesi da considerare è

$$\begin{cases} H_0 : \beta_1 = 0 \\ H_1 : \beta_1 \neq 0 \end{cases}$$

La statistica test da utilizzare è

$\frac{B_1}{\sqrt{\frac{Dev_{disp}(Y)}{(n-2)Dev(X)}}}$ che sotto l'ipotesi H_0 ha distribuzione t_{n-2}. Il valore osservato

della statistica test è

$$t_{oss} = \frac{\hat{\beta}_1}{\sqrt{\frac{Dev_{disp}(Y)}{(n-2)Dev(X)}}} = \frac{2.29}{\sqrt{\frac{17.57}{18 \times 35.41}}} = \frac{2.29}{0.166} = 13.81$$

$$t_{\frac{\alpha}{2};(n-2)} = t_{0.005;18} = 2.8784$$

Si tratta di un test bilaterale, e la regione di rifiuto è $|t_{oss}| > t_{\frac{\alpha}{2};(n-2)} = 2.8784$. Visto che 13.81>2.8784 il test, al prefissato livello di significatività, è significativo. Quindi il coefficiente di regressione osservato non è diverso da zero solo per l'effetto dell'errore di campionamento.

Per quanto riguarda il coefficiente β_0, il sistema di ipotesi da considerare è

$$\begin{cases} H_0 : \beta_0 = 0 \\ H_1 : \beta_0 \neq 0 \end{cases}$$

La statistica test da utilizzare è

$\frac{B_0}{\sqrt{\frac{Dev_{disp}(Y)}{(n-2)}\left[\frac{1}{n} + \frac{\bar{x}^2}{Dev(X)}\right]}}$ che sotto l'ipotesi H_0 ha distribuzione t_{n-2}.

Si tratta anche in questo caso di un test bilaterale, con regione di rifiuto definita da $|t_{oss}| > t_{\frac{\alpha}{2};(n-2)} = 2.8784$.

Il valore osservato della statistica test risulta essere

$$t_{oss} = \frac{\hat{\beta}_0}{\sqrt{\frac{Dev_{disp}(Y)}{(n-2)}\left[\frac{1}{n} + \frac{\bar{x}^2}{Dev(X)}\right]}} = \frac{7.37}{\sqrt{\frac{17.57}{18} \times \left[\frac{1}{20} + \frac{3.41^2}{35.41}\right]}} = \frac{7.37}{0.6069} = 12.1459 > 2.8784$$

Il test è significativo: il valore osservato del coefficiente non è diverso da zero solo per l'effetto dell'errore di campionamento.

Esercizio 13

Su di un campione casuale costituito da 70 individui di sesso maschile in età compresa fra i trenta e i quaranta anni nati nel decennio 1961-1970, è stato rilevato il reddito annuo (Y) in migliaia di euro, e si è osservato:

$$\sum_{i=1}^{70} y_i = 1797.20 \qquad \sum_{i=1}^{70} y_i^2 = 48635.10$$

a) Stimare, giustificando la scelta dello stimatore usato, il reddito medio e la varianza del reddito.
b) Ipotizzando che il reddito abbia distribuzione gaussiana, determinare l'intervallo di confidenza al 90% per il reddito medio.

Soluzione

a) Uno stimatore non distorto per la media di popolazione è la media aritmetica campionaria $\bar{Y} = \frac{1}{n}\sum_{i=1}^{70} Y_i$. In corrispondenza dei dati campionari si ottiene come stima il valore $\bar{y} = \frac{1}{n}\sum_{i=1}^{70} y_i = \frac{1797.10}{70} = 25.67$. Per la varianza di popolazione uno stimatore non distorto è la varianza camponaria corretta $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2$, che in corrispondenza dei dati osservati assume il valore $s^2 = \frac{Dev(Y)}{n-1} = \frac{\sum_{i=1}^{70} y_i^2 - 70\bar{y}^2}{69} = \frac{2493.31}{69} = 35.62$.
b) Per calcolare l'intervallo di confidenza si dovrebbe fare riferimento alla distribuzione t di Student con 69 gradi di libertà, in quanto la varianza non è nota. Se le tavole non forniscono il valore corrispondente, questo può essere approssimato (per difetto) con quello di una t_{70} La numerosità campionaria è comunque abbastanza elevata per fare riferimento anche alla distribuzione Normale. Seguendo questa seconda via, gli estremi dell'intervallo si calcolano quindi usando le espressioni:

$$I_1 = \bar{y} - z_{\alpha/2}\frac{s}{\sqrt{n}} \quad \text{e} \quad I_2 = \bar{y} + z_{\alpha/2}\frac{s}{\sqrt{n}}$$

$1-\alpha = 0.9$ da cui $\frac{\alpha}{2} = 0.05$ e $z_{\frac{\alpha}{2}} = 1.645$; inoltre dal punto a) si ricava $s = \sqrt{35.62} = 5.97$. Gli estremi sono quindi:

$$I_1 = \bar{y} - z_{\alpha/2}\frac{s}{\sqrt{n}} = 25.67 - 1.645 \times \frac{5.97}{\sqrt{70}} = 24.50 \quad I_2 = \bar{y} + z_{\alpha/2}\frac{s}{\sqrt{n}} = 25.67 + 1.645 \times \frac{5.97}{\sqrt{70}} = 26.85$$

Esercizio 14

Con riferimento all'esercizio precedente, supponiamo ora di avere a disposizione per gli stessi individui anche i valori, rilevati congiuntamente, degli anni di istruzione (X). In particolare si ha:

$$\sum_{i=1}^{70} x_i = 1002.33 \qquad \sum_{i=1}^{70} x_i^2 = 15202.52 \qquad \sum_{i=1}^{70} x_i y_i = 27127.11$$

Si vuole studiare la dipendenza lineare del reddito dagli anni di studio, e a questo scopo si considera il modello $Y = \beta_0 + \beta_1 X + \varepsilon$.

a) Stimare, attraverso il metodo dei minimi quadrati, i parametri β_0 e β_1, commentando.
b) Calcolare la devianza di dispersione.
c) Valutare, attraverso un opportuno indice, l'intensità della dipendenza lineare di Y da X, giustificando la scelta dell'indice e commentando il risultato.
d) Secondo la relazione lineare ipotizzata, qual è il reddito che mediamente si dovrebbe associare ad una persona che ha studiato per 13 anni? Commentare.
e) Supponendo che valgano le ipotesi del modello lineare classico, il coefficiente di regressione può essere ritenuto significativamente diverso da zero, ad un livello di significatività del 5%?

Soluzione

a) Le espressioni degli stimatori dei minimi quadrati per i parametri del modello di regressione sono:

$$\hat{\beta}_1 = \frac{Cod(X,Y)}{Dev(X)} \quad \text{e} \quad \hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}.$$

$$\overline{x} = \frac{1002.33}{70} = 14.32 \qquad Dev(X) = \sum_{i=1}^{70} x_i^2 - 70\overline{x}^2 = 850.29$$

$$Cod(X,Y) = \sum_{i=1}^{70} x_i y_i - 70\overline{xy} = 27127.11 - 25733.98 = 1393.13 \text{ e quindi}$$

$$\hat{\beta}_1 = \frac{1393.13}{850.29} = 1.64 \qquad \text{(coefficiente di regressione lineare)}$$

$$\hat{\beta}_0 = 2.21 \qquad \text{(intercetta)}$$

Si osserva la presenza di una relazione di dipendenza lineare diretta di Y da X. Ad un aumento unitario degli anni di istruzione corrisponde un aumento del reddito è pari a 1640 euro.

b) Con i dati che abbiamo a disposizione la devianza di dispersione si calcola attraverso l'uguaglianza

$$Dev_{disp}(Y) = Dev(Y) - Dev_{regr}(Y) = Dev(Y) - \hat{\beta}_1^2 Dev(X) = 2493.31 - 1.64^2 \times 850.29 =$$
$$= 2493.31 - 2282.53 = 210.78$$

c) Per rispondere a questa domanda è necessario calcolare l'*indice di determinazione lineare* $R^2 = \frac{Dev_{regr}(Y)}{Dev(Y)}$.

Tenendo conto delle quantità già calcolate per rispondere ai punti precedenti si ha

$$R^2 = \frac{2282.53}{2493.31} = 0.92$$

Questo indice varia tra 0 (indipendenza lineare perfetta) e 1 (dipendenza lineare perfetta). La dipendenza lineare di Y da X è molto intensa: il 92% della variabilità del reddito è spiegata dalla dipendenza lineare dagli anni di istruzione.

d) Sulla base dei valori stimati per i parametri, il reddito che mediamente risulta associato ad una persona che ha studiato per 13 anni è

$\bar{y} = 2.1 + 1.64 \times 13 = 23.51.$

e) Le ipotesi messe a confronto sono:

$$\begin{cases} H_0 : \beta_1 = 0 \\ H_1 : \beta_1 \neq 0 \end{cases}$$

La statistica test da utilizzare è $t = \dfrac{B_1}{\sqrt{\dfrac{Dev_{disp}(Y)}{(n-2)Dev(X)}}}$ che sotto l'ipotesi nulla ha distribuzione t_{n-2}.

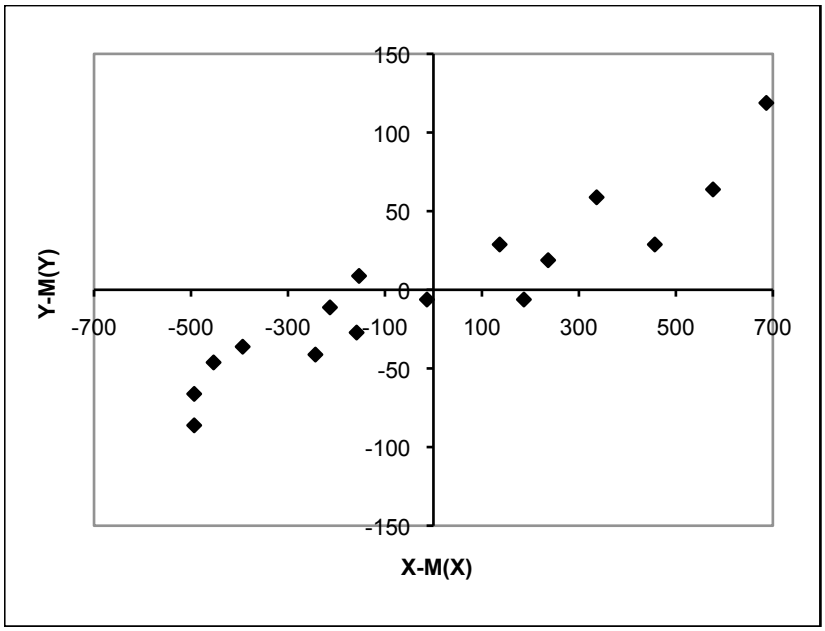
Rifiutiamo H_0 in favore di H_1 se $|t_{oss}| > t_{\frac{\alpha}{2};(n-2)}$, dove $\dfrac{\alpha}{2} = 0.025$. Tenuto conto del fatto che la numerosità campionaria è abbastanza elevata, possiamo approssimare, come nell'esercizio precedente, la distribuzione *t* con la Normale standardizzata. Usando i risultati ottenuti ai punti precedenti si ottiene

$z_{oss} = \dfrac{\hat{\beta}_1}{\sqrt{\dfrac{Dev_{disp}(Y)}{(n-2)Dev(X)}}} = \dfrac{1.64}{\sqrt{\dfrac{210.78}{68 \times 850.29}}} = 27.136$, mentre $z_{0.025} = 1.96$. Il test è significativo e quindi rifiutiamo l'ipotesi H_0: il coefficiente di regressione non differisce da zero per il solo effetto dell'errore di campionamento.

Esercizio 15

Su di un campione costituito da 16 individui sono stati rilevati il reddito netto mensile (X) e la spesa mensile per ristorante/pizzeria (Y). Di seguito sono riportati il diagramma di dispersione delle variabili scarto dalla media X-M(X) e Y-M(Y) e alcune sintesi dei dati.

$\sum_{i=1}^{n} x_i = 31895$ $\sum_{i=1}^{n} y_i = 3539$ $\sum_{i=1}^{n} x_i^2 = 65816625$ $\sum_{i=1}^{n} y_i^2 = 824411$ $\sum_{i=1}^{n} x_i y_i = 7339140$



a) Sulla base del grafico fornire una valutazione qualitativa della connessione fra X e Y.

b) Quantificare l'interdipendenza lineare attraverso il coefficiente di correlazione lineare.

c) Sulla base dei risultati ottenuti, fornire una valutazione della spesa mensile per cene al ristorante di una persona con un reddito netto mensile di 1890 €.

Utilizzando i dati del campione osservato si vogliono stimare i parametri del modello

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

d) Stimare i parametri β_0 e β_1 con gli stimatori dei Minimi Quadrati, commentando il risultato.
e) Stimare la spesa media mensile per cene al ristorante di una persona con un reddito netto mensile pari a 1890 €.
f) Stimare la varianza σ^2 degli errori ε_i, giustificando la scelta dello stimatore usato.
g) Spiegare dettagliatamente la struttura dell'indice di determinazione lineare R^2, calcolarne il valore e interpretare il risultato.

Soluzione

a) I valori sono prevalentemente nel primo e terzo quadrante, e quindi prevalgono gli scarti di segno concorde (+, + oppure -, -). Quindi fra i due caratteri prevale concordanza (a valori più alti/bassi della media di uno si associano valori più alti/bassi della media dell'altro).

b) $\rho = \dfrac{Cod(X,Y)}{\sqrt{Dev(X)Dev(Y)}}$

$$\bar{x} = \frac{31895}{16} = 1993.44 \quad \bar{y} = \frac{3539}{16} = 221.19$$

$$Dev(X) = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2 = 65816625 - 16 \times 221.19^2 = 2235935.94$$

$$Dev(Y) = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2 = 41628.44$$

$$Cod(X,Y) = \sum_{i=1}^{n} x_i y_i - n\overline{xy} = 7339140 - 16 \times 1993.44 \times 221.19 = 284364.69$$

$$\rho = \frac{Cod(X,Y)}{\sqrt{Dev(X)Dev(Y)}} = \frac{284364.69}{\sqrt{2235935 \times 41628.44}} = 0.93$$

Si osserva un'alta correlazione positiva.

c) Il coefficiente di correlazione indica alta correlazione positiva fra i due caratteri (concordanza). Conseguentemente, visto che $\bar{X} = 1993.44$, ci si aspetta che in corrispondenza del valore X=1890, che è più basso della media, si associ un valore di Y anch'esso più basso della media, quindi un valore inferiore a 221.19 €.

d) $\hat{\beta}_1 = \dfrac{Cod(X,Y)}{Dev(X)} = \dfrac{284364.69}{2235935.94} \simeq 0.13$

Il coefficiente di regressione è positivo; all'aumentare del reddito la spesa per cene al ristorante aumenta. Ad esempio, ad un aumento di 100€ del reddito corrisponde un aumento medio della spesa per cene al ristorante di circa 13€.

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = -32.34$$

e) Si usa l'espressione del modello stimato: $\hat{y}_i = -32.34 + 0.13 \times 1890 = 208.03$

f) Uno stimatore non distorto di σ^2 è $s^2 = \frac{Dev_{disp}(Y)}{n-2} = \frac{SQE}{n-2}$

$$Dev_{disp}(Y) = Dev(Y) - Dev_{regr}(Y) = 41628.44 - \hat{\beta}_1^2 Dev(X) =$$
$$= 41628.44 - 36165.29 = 5463.15$$

$$s^2 = \frac{Dev_{disp}(Y)}{n-2} = \frac{5463.15}{14} = 390.23$$

g) $R^2 = \frac{Dev_{regr}}{Dev(Y)} = \frac{36165.29}{41628.44} = 0.8688$

L'intensità del legame di dipendenza lineare di Y da X è elevata: infatti circa l'86.88% della variabilità della spesa per cene al ristorante è spiegata dalla dipendenza lineare dal reddito.

Esercizio 16

Su un campione costituito da 10 supermercati di dimensione simile vengono rilevati superficie (in m^2) dedicata all'esposizione di frutta e verdura provenienti da agricoltura biologica (X) e ricavo (in euro) nella vendita di tali prodotti (Y) in una determinata settimana. I dati sono riportati nella tabella seguente:

X	2.2	1.9	1.6	2.8	1.8	2.3	2.9	2.5	2.0	2.6
Y	501.65	510.80	460.25	620.15	480.00	515.10	570.10	540.25	498.30	585.10

a) Calcolare la superficie media.
b) Calcolare il ricavo mediano.
c) Misurare, attraverso lo scarto quadratico medio, la variabilità nella superficie dedicata all'esposizione.

Sapendo che $\sum_{i=1}^{10} x_i = 22.6 \quad \sum_{i=1}^{10} y_i = 5281.7 \quad \sum_{i=1}^{10} x_i^2 = 52.8 \quad \sum_{i=1}^{10} y_i^2 = 2812242.43$

d) Quantificare l'interdipendenza lineare fra le due variabili.
e) Stimare i parametri della retta di regressione di Y su X, commentando i risultati.
f) Secondo la relazione ipotizzata al punto e), qual è il ricavo che mediamente si dovrebbe associare ad una superficie di esposizione pari a 2.7 m^2 ?
g) Valutare, attraverso un opportuno indice, l'intensità della dipendenza lineare di Y da X, giustificando la scelta dell'indice e commentando il risultato.
h) Supponendo che valgano le ipotesi del modello lineare classico, il coefficiente di regressione può essere ritenuto significativamente diverso da zero, ad un livello di siglificatività dell'1%?

Soluzione

a)La superficie media è pari a $\bar{x} = \frac{1}{10}\sum_{i=1}^{10} x_1 = \frac{22.6}{10} = 2.26$.

b)Per calcolare la mediana è necessario ordinare il carattere. Dalla serie *ordinata* del ricavo

Y	460.25	480.00	498.30	501.65	**510.80**	**515.10**	540.25	570.10	585.10	620.15

si ottengono i due valori centrali **510.80** e **515.10**. Il valore mediano può essere sintetizzato dalla loro media aritmetica, ovvero $\frac{510.8 + 515.1}{2} = 512.95$.

c)

											Tot
X	1.60	1.80	1.90	2.00	2.20	2.30	2.50	2.60	2.80	2.90	**22.60**
X^2	2.56	3.24	3.61	4.00	4.84	5.29	6.25	6.76	7.84	8.41	**52.8**

Abbiamo $Var(X) = \frac{\sum_{i=1}^{10} x_i^2 - 10\overline{x}^2}{10} = \frac{52.8 - 51.08}{10} = \frac{1.72}{10} = 0.17$, da cui

$\sigma(X) = \sqrt{Var(X)} = \sqrt{0.17} = 0.42$.

d) L'interdipendenza lineare può essere quantificata attraverso il Coefficiente di correlazione

$$\rho = \frac{Cod(X,Y)}{\sqrt{Dev(X)Dev(Y)}}$$

Calcoliamo la codevianza:

											Tot
X	1.60	1.80	1.90	2.00	2.2	2.3	2.5	2.6	2.8	2.9	
Y	460.25	480.0	510.80	498.30	501.65	515.1	540.25	585.1	620.15	570.1	
XY	736.40	864.0	970.52	996.60	1103.63	1184.73	1350.62	1521.26	1736.42	1653.29	**12117.48**

$\overline{y} = \frac{1}{10}\sum_{i=1}^{10} y_1 = \frac{5281.7}{10} = 528.17$ e $\overline{x} = \frac{1}{10}\sum_{i=1}^{10} x_1 = \frac{22.6}{10} = 2.26$,

da cui $Cod(X,Y) = \sum_{i=1}^{10} x_i y_i - 10\overline{xy} = 12117.48 - 11936.65 = 180.83$.

$Dev(Y) = \sum_{i=1}^{10} y_i^2 - 10\overline{y}^2 = 2812242.43\text{-}10 \times 528.17^2 = 22606.94$

$Dev(X) = \sum_{i=1}^{10} x_i^2 - 10\overline{x}^2 = 52.80\text{-}10 \times 2.26^2 = 1.72$

$$\rho = \frac{Cod(X,Y)}{\sqrt{Dev(X)Dev(Y)}} = \frac{180.83}{\sqrt{1.72 \times 22606.94}} = 0.92$$

Le due variabili sono concordi (il segno del coefficiente di correlazione è positivo), e si riscontra una intensa interdipendenza lineare, visto che $-1 \le \rho \le 1$, e il valore osservato è molto vicino a 1.

e) Stimiamo i parametri del modello di regressione con le espressioni relative al metodo dei minimi quadrati:

$\hat{\beta}_1 = \frac{Cod(X,Y)}{Dev(X)} = \frac{180.83}{1.72} = 104.89$ (coefficiente di regressione lineare)

$\hat{\beta}_0 = \overline{y} - b\overline{x} = 528.17 - 237.05 = 291.12$ (intercetta)

Si osserva la presenza di una relazione di dipendenza lineare diretta di Y da X. Ad un aumento unitario della superficie dedicata all'esposizione corrisponde un aumento del ricavo settimanale pari a 104.89 euro.

f) Sulla base della relazione lineare stimata si ottiene
$\overline{y} = 291.12 + 104.89 \times 2.7 = 574.32$

g) Per rispondere a questa domanda è necessario calcolare l'*indice di determinazione lineare* $R^2 = \frac{Dev_{regr}(Y)}{Dev(Y)}$.

$Dev(Y) = \sum_{i=1}^{10} y_i^2 - 10\bar{y}^2 = 2812242.43 - 2789635.49 = 22606.94$

Tenendo conto delle quantità già calcolate per rispondere ai punti precedenti, per calcolare il numeratore è conveniente usare l'espressione

$Dev_{regr}(Y) = \hat{\beta}_1^2 Dev(X) = 18967.85$, e quindi

$R^2 = \frac{18967.85}{22606.94} = 0.83$

Questo indice varia tra 0 (indipendenza lineare perfetta) e 1 (dipendenza lineare perfetta). La dipendenza lineare di Y da X è molto intensa: l'83% della variabilità del ricavo è spiegata dalla dipendenza lineare dalla superficie dedicata all'esposizione.

h)Le ipotesi messe a confronto sono:

$\begin{cases} H_0 : \beta_1 = 0 \\ H_1 : \beta_1 \neq 0 \end{cases}$

La statistica test da utilizzare è $t = \frac{B_1}{\sqrt{\frac{Dev_{disp}(Y)}{(n-2)Dev(X)}}}$ che sotto l'ipotesi nulla ha distribuzione t_{n-2}.

Rifiutiamo H_0 in favore di H_1 se $|t_{oss}| > t_{\frac{\alpha}{2};(n-2)}$. Usando i risultati ottenuti ai punti precedenti si ottiene

$t_{oss} = \frac{\beta_1}{\sqrt{\frac{Dev_{disp}(Y)}{(n-2)Dev(X)}}} = \frac{104.89}{\sqrt{\frac{3639.09}{8 \times 1.72}}} = 6.46$, mentre $t_{0.005;8} = 3.3554$. Al prefissato livello di significatività il test è significativo: il coefficiente di regressione stimato non differisce da zero per il solo effetto dell'errore di campionamento.