Politecnico di Milano – Facoltà di Ingegneria Informatica

Anno Accademico 2008/2009

Corso di Statistica (2L) per INF e TEL

Docente: Antonio Pievatolo; esercitazioni: Raffaele Argiento

Esercitazione del 22/05/2009

Esercizio 1(dal Tema d'esame del corso Statistica Matematica del 25/02/2008)

Viene condotto uno studio per capire se seguire un corso di statistica faccia aumentare il quoziente di intelligenza (IQ). Per tale motivo vengono scelti a caso 12 studenti universitari tra quelli iscritti al corso e si registrano i loro IQ prima e dopo avere seguito il corso di statistica. I dati ottenuti sono i seguenti

IQ prima	104	125	127	102	140	122	118	110	126	138	116	125
IQ dopo	111	120	138	113	142	130	114	121	135	145	118	125.

Si assuma che i quozienti di intelligenza e la loro differenza siano variabili aleatorie gaussiane.

1. Si stabilisca se, in questo caso, esiste una chiara evidenza sperimentale che seguire il corso di statistica faccia aumentare il quoziente di intelligenza.

2. Si calcoli un intervallo di confidenza bilaterale di livello 99% per la differenza delle medie dei due quozienti intellettivi prima e dopo avere seguito il corso.

Soluzione

Sia $X_P \sim N(\mu_P, \sigma_P^2)$ la variabile aleatoria che rappresenta l'IQ di uno studente universitario prima di avere seguito il corso di statistica e $X_D \sim N(\mu_D, \sigma_D^2)$ quella che rappresenta l'IQ dello stesso studente dopo avere seguito il corso. I dati (X_P, X_D) sono accoppiati.

1. Si vuole studiare il test $H_0 : \ \mu_P \geq \mu_D$ contro $H_1 : \mu_P < \mu_D$. Allora si considera la differenza $W := X_P - X_D$ da un campione normale con dati osservati

   $$(-7, 5, -11, -11, -2, -8, 4, -11, -9, -7, -2, 0).$$

   Si tratta di verificare $H_0 : \ \mu_W \geq 0$ contro $H_1 : \mu_W < 0$ e la statistica test è

   $$U := \frac{\bar{W}}{\sqrt{S_W^2/12}},$$

   che ha distribuzione T_{11} se H_0 è vera. Da $\bar{w} = -4.9167$, $s_W^2 = 33.1742$, si ottiene $u = -2.9571$. Il il p-value è $P(T_11 < -2.9571) = P(T_11 > 2.9571) = 0.0065$ [con il solo uso delle tavole si ottiene $0.005 <$ p-value < 0.01]; pertanto si rifiuta l'ipotesi nulla al livello $\alpha = 0.01$, inoltre si conclude che esiste una forte evidenza sperimentale a favore di H_1.

2. Si osservi che

   $$\frac{\bar{W} - \mu_W}{\sqrt{S_W^2/n}} \sim T_{n-1}.$$

   Quindi, l'intervallo di confidenza bilaterale di livello $(1-\alpha)100\%$ per $\mu_P - \mu_D = \mu_W$ è individuato dai punti $\bar{w} \pm t_{\alpha/2,n-1} \frac{s_W}{sqrtn}$, con $t_{\alpha/2,n-1} = t_{0.005,11} = 3.106$; pertanto si trova

   $$(-10.0810, 0.2476).$$

∎

Esercizio2

Un investitore finaziario vuole verificare se c'è dipendenza fra la variazione in percentuale dell'indice Down Jones nei primi 5 giorni di contrattazione (Variabile X) e la variazione percentuale dell'indice in tutto l'anno (variabile Y). Per $n = 13$ anni consecutivi , l'investitore registra la variazione percentuale del Down Jones, ottenendo la seguente una realizzazione campionaria che riassumiamo con le seguenti statistiche:

$$\bar{x} = 0.5538, \quad \bar{y} = 11.8153, \quad \sum_{i=1}^{n} = x_i^2 = 80.06, \quad \sum_{i=1}^{n} y_i^2 = 3718.76, \quad \sum_{i=1}^{n} x_i y_i = -69.67.$$

Sotto l'ipotesi che X e Y siano distribuite normalmente

1. Fornite una stima del coefficiente di correlazione lineare fra X e Y.
2. Verificate che i dati NON evidenziano dipendenza fra la variazione del Down Jones nei primi giorni dell'anno e la variazione annuale ($\alpha = 0.1$).

Soluzione

1. Il coefficiente di correlazione lineare è definito definito da:

$$\rho := \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}}.$$

Una stima ρ è data dal coefficiente di correlazione lineare campionario:

$$R = \frac{\mathrm{Cov}_{X,Y}}{S_X^2 S_Y^2} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \times \sum_{i=1}^{n}(Y_i - \bar{Y})}}.$$

Dove $\mathrm{Cov}_{X,Y}$ è la covarianza campionaria fra X e Y definita da:

$$\mathrm{Cov}_{X,Y} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}) = \frac{n}{n-1}\left(\frac{1}{n}\sum_{i=1}^{n} X_i Y_i - \bar{X}\bar{Y}\right).$$

Ricordiamo ora che:

$$S_X^2 = \frac{n}{n-1}\left(\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}^2\right) \text{ e } S_Y^2 = \frac{n}{n-1}\left(\frac{1}{n}\sum_{i=1}^{n} Y_i^2 - \bar{Y}^2\right).$$

Dunque tramite valori selle statistiche riportati dal testo dell'esercizio, ricaviamo le realizzazioni:

$$s_X^2 = 6.339 \quad s_Y^2 = 158.6619, \quad cov_{X,Y} = -12.8944.$$

Concludiamo calcolando $r = \frac{c_{X,Y}}{\sqrt{s_X^2 \times s_Y^2}} = -0.4066$

2. La coppia (X, Y) è congiuntamente gaussiana, dunque X e Y sono indipendenti se e soltanto se $\rho = 0$, dunque studiamo il test:

$$\mathrm{H}_0\colon\ \rho = 0 \text{ contro } \mathrm{H}_1\colon\ \rho \neq 0.$$

Ricordiamo che se sotto H_0 la statistica

$$U := \frac{R}{\sqrt{R^2}}\sqrt{n-1}, \quad n \geq 3$$

ha distribuzione t-student con $n - 1$ gradi di libertà. Il valore osservato di U è $u = -1.4760$ e dunque $\frac{\text{p-value}}{2} = \mathbb{P}(U \leq u) = 1 - \mathbb{P}(U \leq 1.4760)$. Dalle tavole ricaviamo che $0.1 <$ p-value < 0.2. Se $\alpha = 0.1$ allora $\alpha <$p-value, quindi NON rifiuto H_0.

∎

Esercizio 3

Una grossa industria vuole confrontare le caratteristiche tra due resine plastiche per costruire tubature. La resina 1 è meno costosa della resina 2, ma si sospetta che la resina 2 abbia un maggior allungamento medio percentuale (capacità del materiale di allungarsi senza superare la rottura). Un campione sperimentale di dimensione 120 della resina 1 ha fornito un'allungamento medio percentuale pari a 2.1 e una deviazione standard campionaria pari a 0.51, mentre un campione di dimensione 130 della resina 2 ha fornito un'allungamento medio percentuale pari a 2.5 e una deviazione standard campionaria pari a 0.58. Si assuma che le varianze effettive delle due resine siano uguali.

1. Si fornisca una stima della varianza σ^2 delle due resine.
2. I dati a disposizione permettono di concludere a livello 1% che l'allungamento medio della resina 2 è maggiore di quello della resina 1? Si costruisca un opportuno test statistico.
3. Si costruisca un intervallo di confidenza di livello 98% per la differenza tra gli allungamenti medi delle due resine considerate.
4. Si ripeta il test al punto b) senza l'ipotesi che le varianze effettive delle due resine siano uguali.

Soluzione

1. $S_P^2 = \dfrac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2} = 0.30.$
2. $H_0 : \mu_1 = \mu_2$ contro $H_1 : \mu_2 > \mu_1$.
   Statistica test:
   $$Z_0 = \frac{\bar{x}_2 - \bar{x}_1}{S_P\sqrt{1/n_1 + 1/n_2}},$$
   da cui $z_0 = 5.77 > z_\alpha = 2.33$; in conclusione rigetto H_0.
3. $IC_{98\%}(\mu_2 - \mu_1) = \left(\bar{x}_2 - \bar{x}_1 \pm z_{\alpha/2} S_P\sqrt{1/n_1 + 1/n_2}\right) = (0.4 \pm 0.16) = (0.24; 0.56).$
4. La varianza campionaria S_1^2 è uno stimatore di σ_1^2 e S_2^2 è uno stimatore di σ_2^2. Inoltre dato che $n_1 > 50$ e $n_2 > 50$ si ha che approssimativamente
   $$\frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \overset{H0}{\sim} N(0,1).$$
   Il valore osservato dalla statistica test è
   $$z = \frac{-0.4}{\sqrt{\frac{0.51^2}{120} + \frac{0.58^2}{130}}} = -5.80.$$
   Dunque, indicando con Z una variabile con distribuzione normale standard, p-value=$P(Z < -5.80) = 3.3 * 10^{-9}$. C'è dunque una forte evidenza sperimentale contro l'ipotesi nulla.

∎