Politecnico di Milano - Scuola di Ingegneria Industriale e dell'Informazione

Appello di Statistica per Ingegneria Energetica
24 luglio 2014

Cognome, Nome e Numero di matricola:

Problema 1. Leonard Leakey Hofstadter considera l'isocronismo del pendolo semplice per oscillazioni di piccola ampiezza descritto da Galileo: a prescindere dalla loro ampiezza, le oscillazioni di piccola ampiezza si svolgano tutte nello stesso tempo t (detto periodo), dipendente solo dalla lunghezza del pendolo.

Tuttavia, quando uno studente del Caltech deve verificare sperimentalmente tale legge con il pendolo dell'Istituto, ad ogni misura ottiene un risultato T affetto da errore. Leonard suppone $T \sim N(t, \sigma^2)$.

1. Sapendo che il 5% delle misure dà valori inferiori o ugali a 1.515 e che l'8% delle misure dà valori maggiori o uguali a 1.542, quanto valgono la media t e la varianza σ^2 di T.
2. Calcolare la probabilità che una singola misura dia un risultato $T > t$.
3. Calcolare la probabilità che su 120 misure almeno 72 diano un risultato $T_k > t$.

Risultati.

1. $$P[T \le 1.515] = 0.05; \qquad P[T \ge 1.542] = 0.08$$

$$P[\frac{T-t}{\sigma} \le \frac{1.515-t}{\sigma}] = 0.05; \qquad P[\frac{T-t}{\sigma} \ge \frac{1.542-t}{\sigma}] = 0.08 \implies P[\frac{T-t}{\sigma} \le \frac{1.542-t}{\sigma}] = 0.92$$

Detta Φ la funzione di ripartizione della $N(0;1)$ abbiamo il sistema:

$$\begin{cases} \dfrac{1.515-t}{\sigma} = \Phi^{-1}(0.05) = -\Phi^{-1}(0.95) = -1.6449 \\ \dfrac{1.542-t}{\sigma} = \Phi^{-1}(0.92) = 1.4051 \end{cases}$$

$$\implies \qquad t = 1.5296, \qquad \sigma = 0.00885, \qquad \sigma^2 = 7.832 \cdot 10^{-5}$$

2. $P(T > t) = 0.5$.
3. Costruiamo 120 nuove variabili aleatorie W_i di Bernoulli così definite:

$$W_i = \begin{cases} 1 & \text{se } T_i > t \\ 0 & \text{se } T_i \le t \end{cases} \implies P[W_i = 1] = 0.5; \quad P[W_i = 0] = 0.5$$

$$Y = \sum_{i=1}^{120} W_i \sim B(120; 0.5)$$

Le condizioni per l'approssimazione gaussiana via Teorema Centrale sono soddisfatte. Dunque:

$$Y = \sum_{i=1}^{120} W_i \approx N(120 \cdot 0.5; 120 \cdot 0.5^2) = N(60; 30)$$

$$P(Y \ge 72) = P(Y > 71.5) = P\left(\frac{Y-60}{\sqrt{30}} > \frac{71.5-60}{\sqrt{30}}\right) \simeq 1 - \Phi(2.10) = 0.018.$$

Problema 2. L'ingegnere aerospaziale Howard Wolowitz, del California Institute of Technology, è chiamato, dopo il suo soggiorno sulla Stazione Spaziale Internazionale, a svolgere alcune analisi sul numero X di orbite giornaliere compiute dalla stazione stessa. In particolare si vuole confrontare X col numero Y di orbite giornaliere compiute da una nuova stazione spaziale, da poco operante nell'orbita terrestre, per capire se questa compia in media lo stesso numero di orbite giornaliere.

Per la Stazione Spaziale Internazionale si intende quindi osservare un campione $X_1, \dots, X_{n_X}$ di orbite giornaliere per n_X giorni consecutivi. Analogamente, per la nuova stazione spaziale si intende osservare un campione $Y_1, \dots, Y_{n_Y}$ di orbite giornaliere per n_Y giorni consecutivi.

Si supponga che X e Y siano variabili aleatorie normali a media incognita e varianza nota, pari a 2 per la Stazione Spaziale Internazionale e a 1.8 per la nuova stazione spaziale. Si supponga anche che i suddetti campioni risultino casuali e indipendenti fra di loro.

Per ciascuna stazione

a) introdurre uno stimatore corretto del numero medio di orbite giornaliere e calcolarne l'errore quadratico medio;

b) determinare il numero di giorni minimo durante i quali devono essere raccolti i dati affinché i due stimatori abbiano un errore quadratico medio non superiore a 0.1.

A questo punto vengono messi a disposizione di Wolowitz i dati relativi al numero di orbite giornaliere compiute dalle due stazioni spaziali durante il mese di aprile 2014, ottenendo una media campionaria pari a 15.7 per la Stazione Spaziale Internazionale e una media campionaria pari a 16 per la nuova stazione spaziale.

c) Costruire un intervallo di confidenza bilatero, al 95%, per la differenza tra le due medie.

d) Si consideri un test di significatività all'1% per testare l'ipotesi che le due medie siano diverse. Indicare, senza ulteriori calcoli e con opportune giustificazioni, la risposta del test. Si tratta di una conclusione forte o debole?

e) Calcolare il p-value dei dati relativo al test precedente.

Risultati.

a) $\widehat{\mu}_X = \overline{X} \quad \Rightarrow \quad \mathrm{MSE}(\overline{X}) = \sigma_X^2/n_X = 2/n_X$
$\widehat{\mu}_Y = \overline{Y} \quad \Rightarrow \quad \mathrm{MSE}(\overline{Y}) = \sigma_Y^2/n_Y = 1.8/n_Y.$

b) $MSE(\overline{X}) = \sigma_X^2/n_X = 2/n_X \leq 0.1 \to n_X \geq 20;$
$MSE(\overline{Y}) = \sigma_Y^2/n_Y = 1.8/n_Y \leq 0.1 \to n_Y \geq 18.$

c) Siano ora $n_X = n_Y = 30$ e $\gamma = 0.95$. Ricaviamo $IC_\gamma(\mu_X - \mu_Y) = \overline{x} - \overline{y} \pm z_{\frac{1-\gamma}{2}} \sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}} = -0.3 \pm 0.7 = [-1; 0.4]$.

d) Dato che lo zero è contenuto nell'intervallo di confidenza al 95%, significa che con un test al 5% non potremmo rifiutare l'ipotesi nulla di uguaglianza fra le due medie. A maggior ragione non possiamo rifiutare neanche ad un livello di significatività dell'1%. Si tratta di una conclusione debole.

e) $z_0 = \frac{\overline{x} - \overline{y}}{\sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}}} = -0.84$, quindi p-value $= 2\,\mathbb{P}(Z > |z_0|) = 0.4$.

Problema 3. L'astrofisico Raj Koothrappali ha raccolto per mesi dati relativi ad alcune stelle della Via Lattea. Il motivo del suo meticoloso studio è valutare cosa caratterizza la luminosità di una stella nella nostra galassia. Il dataset che è riuscito a comporre consta alla fine di 43 osservazioni in cui vengono registrate per ogni stella la luminosità, la temperatura e la costante di riflettività.

L'astrofisico elabora un primo modello in cui cerca di spiegare la luminosità utilizzando entrambe le covariate a disposizione e la loro interazione, ottenendo il seguente output:

```
lm(formula = light ~ temp + rifl + temp * rifl, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-0.84543 -0.32884 -0.02732  0.28389  0.88119

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -4.5475     2.0965  -2.169   0.0362 *
temp          2.1580     0.4777   4.518 5.66e-05 ***
rifl          1.0406     1.9944   0.522   0.6048
temp:rifl    -0.2300     0.4563  -0.504   0.6170
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.4126 on 39 degrees of freedom
Multiple R-squared:  0.377,        Adjusted R-squared:  0.3291
F-statistic: 7.868 on 3 and 39 DF,  p-value: 0.0003181
```

Dall'osservazione dell'output, Raj si rende conto subito che il modello è ridondante, e prova quindi ad eliminare le variabili che non risultano significative, ottenendo:

```
lm(formula = light ~ temp, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-0.8097 -0.3088 -0.0267  0.2866  0.9078

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -4.0565     1.8441  -2.200   0.0335 *
temp          2.0467     0.4202   4.871  1.7e-05 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.4058 on 41 degrees of freedom
Multiple R-squared:  0.3666,        Adjusted R-squared:  0.3511
F-statistic: 23.73 on 1 and 41 DF,  p-value: 1.697e-05
```

Per entrambi i modelli Raj non ha motivo di dubitare della bontà dell'ipotesi gaussiana e, confortato da questo nuovo risultato, sarebbe intenzionato a scegliere il secondo modello. Tuttavia, prima di prendere una decisione definitiva, vorrebbe verificare con un opportuno test statistico che l'eliminazione delle variabili non comporti una perdita significative della capacità predittiva. Imposta quindi un test di confronto tra i due modelli, ottenendo un p-value pari a 0.72, ma non sa come leggerlo e che conclusioni trarre.

a) Esplicitare il test da svolgere, indicando ipotesi nulla, ipotesi alternativa, regione critica.

b) Indicare a Raj la corretta interpretazione del p-value.

Convintosi della bontà del suo ragionamento, Raj vuole ora utilizzare il modello ridotto per fare inferenza sulla luminosità delle stelle in funzione della loro temperatura. Come prima cosa, vuole ricavare una

relazione immediata per capire di quanto ci si può aspettare che incrementi la luminosità di una stella con il crescere della sua temperatura.

c) Stimare il tasso di incremento medio nella luminosità di una stella al crescere di un grado di temperatura.

Raj vuole essere ulteriormente sicuro in relazione ai conti che ha fatto, pertanto vorrebbe accertarsi, con un test all'1% di significatività, che tale tasso sia maggiore di 1.

d) Impostare un test opportuno che consenta a Raj di verificare la sua congettura, definendo ipotesi nulla, ipotesi alternativa, regione critica, quindi rispondere all'astrofisico.

e) Calcolare l'intervallo di previsione al 95% per la luminosità una stella di temperatura pari a 4, sapendo che per i dati raccolti la media delle temperture ($\bar{x}$) è pari a 4.386 e la somma degli scarti dalla media al quadrato (S_{xx}) è pari a 0.933.

Risultati.

a) Deatta Y la luminosità, x_1 la temperatura e x_2 la costante di riflettività, si parte dal modello empirico gaussiano completo:

$$Y = \beta_0 + \beta_1\, x_1 + \beta_2\, x_2 + \beta_{12} x_1\, x_2 + \epsilon, \qquad \epsilon \sim N(0, \sigma^2).$$

Il test F da impostre è il seguente:

$$H_0 \ : \ \beta_2 = \beta_{12} = 0 \qquad \text{vs} \qquad H_1 \ : \ \beta_2 \neq 0 \text{ o } \beta_{12} \neq 0$$

$$R_\alpha = \{F_0 > f_\alpha^{3-1,43-4}\}, \qquad F_0 = \frac{(SSE_{red} - SSE_{tot})/(3-1)}{SSE_{tot}/(43-4)}$$

dove $f_\alpha^{2,39}$ è il punto percentuale di ordine α di una Fisher a 2 e 39 gradi di libertà.

b) Il p-value alto ($>> 5\%$) indica che non vi è evidenza per ritenere che il modello ridotto sia meno valido di quello completo, in termini di capacità previsiva.

c) Quanto richiesto corrisponde alla stima del coefficiente β_1 del modello di regressione scelto, ovvero $\widehat{\beta}_1 = 2.047$.

d) $H_0 \ : \ \beta_1 \leq 1$ vs $H_0 \ : \ \beta_1 > 1$. Statistica test: $T_0 = \frac{\hat{\beta}_1 - 1}{se(\hat{\beta}_1)} = 2.492$. Regione critica: $R_\alpha = \left\{T_0 > t_\alpha^{n-1}\right\}$, dove $\alpha = 0.01$. Si ottiene pertanto $T_0 > t_\alpha^{(42)}$, quindi Raj può rifiutare l'ipotesi nulla e affermare con la confidenza richiesta che β_1 è maggiore di 1.

e) Sfruttando la stima della retta $-4.0565 + 2.0467 * 4$ e di $\sigma^2 = 0.4058^2$ dall'output di R, sapendo che il quantile di ordine $(1-\alpha/2)$ di una t a $(43-2=41)$ gdl vale 2.019, e sfruttndo le informazioni date, si ottiene

$$IP_{0.95}(y_4) = -4.0565 + 2.0467 * 4 \pm 2.019 * \sqrt{0.4058^2 * (1 + \frac{1}{43} + \frac{0.386^2}{0.933})}$$
$$= 4.1303 \pm 0.9308 = (3.1995\,,\, 5.0611)$$