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

II PROVA IN ITINERE DI STATISTICA PER INGEGNERIA ENERGETICA
4 luglio 2013

Cognome, Nome e Numero di matricola:

Problema 1. Supponendo di non conoscere alcuna cifra decimale di π, il matematico Mario Lazzarini vuole stimarne statisticamente il valore lanciando n aghi di lunghezza ℓ su un pavimento a strisce distanti $6\ell/5$, cosicché la probabilità che un singolo ago intersechi una qualche linea del pavimento valga

$$p = \frac{5}{3\pi}.$$

Sia X_k la variabile che indica se il k-esimo ago interseca una linea del pavimento e sia $S = \sum_{k=1}^{n} X_k$ il numero totale di aghi lanciati che intersecano linee del pavimento.

(a) Come sono distribuite le variabili X_k ed S?

(b) Proporre uno stimatore $\widehat{p}$ per p e dedurne uno stimatore $\widehat{\pi}$ per π.

(c) Proporre un intervallo di confidenza di livello 0.9 per p e dedurne un intervallo di pari livello per π.

(d) Quanti aghi dovrebbe lanciare il Lazzarini per ottenere un intervallo per p lungo al massimo 0.03?

Il Lazzarini infine lancia 3˙408 aghi contando 1˙808 intersezioni.

(e) Fornire una stima puntuale di π (tenendo almeno 7 cifre significative dopo la virgola!).

(f) Fornire una stima intervallare di π di livello 0.9.

(g) Supponendo che π appartenga effettivamente all'intervallo trovato in (f), quale errore commetto al massimo se uso per π il valore trovato in (e)?

Soluzione.

(a) $X_k \sim B(p)$ per ogni $k = 1, \dots, n$, mentre $S \sim B(n, p)$.

(b) $\widehat{p} = \overline{X}_n = \dfrac{S}{n}$ da cui $\widehat{\pi} = \dfrac{5}{3\widehat{p}} = \dfrac{5n}{3S}$.

(c) IC per p:

$$\widehat{p} - \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}1.645 < p = \frac{5}{3\pi} < \widehat{p} + \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}1.645$$

da cui IC per π:

$$\frac{5}{3\left(\widehat{p} + \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}1.645\right)} < \pi < \frac{5}{3\left(\widehat{p} - \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}1.645\right)}$$

(d) $2\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}1.645 \le \frac{1.645}{\sqrt{n}} \le 0.03$ da cui $n \ge \left(\frac{1.645}{0.03}\right)^2 = 3\dot{}006.694$ cioè $n \ge 3\dot{}007$.

(e) $\pi \simeq \widehat{\pi} = \frac{5}{3}\,\frac{3\dot{}408}{1\dot{}808} = 3.1415929$.

(f) $3.0604660 < \pi < 3.2271379$.

(g) L'errore massimo è il massimo fra $|3.1415929 - 3.0604660|$ e $|3.1415929 - 3.2271379|$ cioè 0.085545.

Problema 2. Uno studente dell'ultimo anno di scuola superiore, intendendo iscriversi ad un corso di Laurea in Ingegneria ed essendo indeciso tra il Politecnico di Torino ed il Politecnico di Milano, ha negli ultimi giorni monitorato il numero di visite giornaliere ai relativi siti, registrando i seguenti valori.
Politecnico di Torino: 2270 2490 2190 2030 2540 2710
Politecnico di Milano: 2740 3600 3120 3110 3280 3280 2340 2460
Nel seguito si assuma che il numero di visite giornaliere al sito del Politecnico di Milano e il numero di visite giornaliere al sito del Politecnico di Torino siano indipendenti ed entrambi normalmente distribuiti.

1. Verificare con un opportuno test con un livello di significatività pari a 5% se le varianze delle due popolazioni possono essere considerate uguali, calcolandone il p-value.

2. Basandosi sul risultato del test al punto precedente, calcolare un intervallo di confidenza di livello 95% per la differenza delle medie del numero di visite giornaliere dei siti del Politecnico di Torino e del Politecnico di Milano.

3. E' possibile affermare ad un livello di significatività del 5% che le medie del numero di visite giornaliere dei due siti sono diversi? Stimare il p-value.

Soluzione.

1. I dati possono essere considerati normali: $X_T \sim N(\mu_T, \sigma_T^2)$, $X_M \sim N(\mu_M, \sigma_M^2)$. Si vuole effettuare il test $H_0 : \sigma_T = \sigma_M$ vs. $H_1 : \sigma_T \neq \sigma_M$. La regione critica di questo test è

$$R = \left\{(x_1^L, \ldots, x_{n_T}^L), (x_1^P, \ldots, x_{n_M}^P) : \frac{s_T^2}{s_M^2} > f_{\alpha/2, n_T-1, n_M-1} \text{ or } \frac{s_T^2}{s_M^2} < f_{1-\alpha/2, n_T-1, n_M-1}\right\}.$$

   Dato che varianze campionarie sono $s_T^2 = 63376.67$ e $s_M^2 = 191069.6$, si ottiene $f_0 = s_T^2/s_M^2 = 0.3317$. Poichè

$$f_{0.975,5,7} = 1/f_{0.025,7,5} = 1/6.85 = 0.146 < f_0 = 0.3317 < f_{0.025,5,7} = 5.29$$

   non si può rifiutare l'ipotesi nulla ad un livello del 5%.

   Poichè per $\alpha = 0.2$

$$f_{0.9,5,7} = 1/f_{0.1,7,5} = 1/3.37 = 0.30 < f_0 = 0.3317 < f_{0.1,5,7} = 2.88$$

   mentre per $\alpha = 0.5$

$$f_{0.75,5,7} = 1/f_{0.25,7,5} = 1/1.89 = 0.53 \not< f_0 = 0.3317 < f_{0.25,5,7} = 1.71$$

   otteniamo 20% < p-value < 50%.

2. Dato che le varianze delle due popolazioni possono essere considerate uguali e i dati sono non accoppiati, l'intervallo di confidenza di livello 95% per la differenza di medie è

$$IC_{0.95} = \left(\bar{x}_T - \bar{x}_M \pm t_{\alpha/2, n_T+n_M-2} s_M \sqrt{\frac{1}{n_T} + \frac{1}{n_M}}\right)$$

   La varianza pooled è pari $s_M^2 = \frac{(n_T-1)s_T^2 + (n_M-1)s_M^2}{n_T+n_M-2} = 137864.2$, $s_M\sqrt{\frac{1}{n_T} + \frac{1}{n_M}} = 200.52$ e $t_{0.025,12} = 2.179$ per cui si ottiene

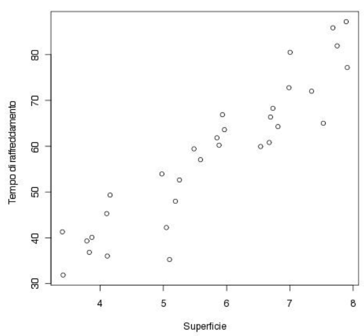
$$IC_{0.95} = (-619.5833 \pm 436.9071) = (-1056.49, -182.6762)$$

3. Dato che l'intervallo di confidenza calcolato al punto precedente non contiene il valore 0, si può affermare (ad un livello di significatività del 5%) che il numero di visite giornaliere ai due siti sia significativamente differente. Il p-value è 0.009366.

Problema 3. Per i cilindri metallici prodotti dalla ABC, aventi tutti la stessa lunghezza, ma differente raggio, vogliamo studiare il tempo di raffreddamento Y (inteso come i minuti impiegati a tornare alla temperatura di 20^oC dopo essere stati riscaldati a 50^oC) in relazione alla superficie laterale x (variabile tra 3 e 8 cm^2). Assumiamo valido il modello gaussiano empirico

$$Y = \beta_0 + \beta_1 x + \epsilon, \qquad \epsilon \sim N(0, \sigma^2).$$

I dati relativi ad uno stock di $n = 32$ cilindri sono stati elaborati con R, ottenendo il *diagramma di dispersione* e l'output della regressione lineare qua riportati:




```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.6920     4.5288  -0.153     0.88
superficie   10.2198     0.7635  13.385 3.47e-14

Residual standard error:  6.031 on 30 degrees of freedom
Multiple R-squared:  0.8566, Adjusted R-squared:  0.8518
F-statistic:  179.2 on 1 and 30 DF, p-value:  3.474e-14
```

a) Fornire una stima puntuale per la varianza degli errori ϵ.

b) Fornire una previsione intervallare al 90% per il tempo di raffreddamento di una 33-esima sbarretta avente superficie laterale pari a 9 cm^2.

c) Verificare se possiamo ritenere $\beta_0 = 0$ mediante un opportuno test di ipotesi, specificando: ipotesi nulla, ipotesi alternativa, regione critica di livello α, p-value dei dati raccolti, conclusione.

d) Verificare se possiamo ritenere $\beta_1 = 8$ mediante un opportuno test di ipotesi, specificando: ipotesi nulla, ipotesi alternativa, regione critica di livello α, p-value dei dati raccolti, conclusione.

Soluzioni.

a) La stima puntuale della deviazione standard degli errori è $\hat{\sigma} = 6.031$, pertanto

$$\hat{\sigma}^2 = 6.031^2 = 36.372961.$$

b) Al 90% in corrispondenza di $x = 9$, usando la formula nota si ottiene $Y(9) \in [80.1, 102.5]$.

c) $H_0 : \beta_0 = 0$, $H_1 : \beta_0 \neq 0$, $R_\alpha : |\widehat{\beta}_0| > \text{se}(\widehat{\beta}_0) t_{\alpha/2}(30)$,
p-value $= 0.88$, conclusione (debole): $\beta_0 = 0$.

d) $H_1 : \beta_1 = 8$, $H_1 : \beta_1 \neq 8$, $R_\alpha : |\widehat{\beta}_1 - 8| > \text{se}(\widehat{\beta}_1) t_{\alpha/2}(30)$,
$0.005 <$ p-value $= 0.0067939318 < 0.01$,
conclusione (forte) agli usuali livelli di significatività: $\beta_1 \neq 8$.