Politecnico di Milano - Scuola di Ingegneria Industriale

II Prova in Itinere di Statistica per Ingegneria Energetica
7 Luglio 2011

Cognome, Nome e Numero di matricola:

Problema 1. La società Hastings s.r.l. viene incaricata di svolgere un'indagine sull'altezza degli studenti del Politecnico. Viene misurata l'altezza X di 120 studenti e una sintesi dei dati (misurati in metri) è riportata in tabella:

Classi	**Freq. Assoluta**
$X < 1.60$	10
$1.60 \leq X < 1.75$	49
$1.75 \leq X < 1.90$	36
$X \geq 1.90$	25

L'attenzione della ricerca si concentra sulla proporzione p di studenti con altezza almeno pari a 1.75 m (cioè $p = P[X \geq 1.75]$).

(a) Si fornisca una stima puntuale per p.

(b) Si dichiari la regione critica di un test di livello $\alpha = 0.05$ per

$$H0 : p = 0.4 \quad \text{v.s.} \quad H1 : p \neq 0.4$$

(c) Si calcoli il p-value del test di cui al punto precedente e si dica a quale conclusione porta quindi il test.

(d) Nel caso si venisse a sapere di aver preso la decisione sbagliata nel test al punto (b), che tipo di errore si sarebbe commesso?

(e) Calcolare l'errore di II tipo per il test effettuato al punto (b), nel caso la proporzione vera fosse $p = 0.5$.

Soluzione.
(a) Le osservazioni possono essere descritte come variabili bernoulliane

$$A_i = \begin{cases} 1 & \text{se} \quad X_i \geq 1.75 \\ 0 & \text{se} \quad X_i < 1.75 \end{cases},$$

la cui media è esattamente il parametro p che vogliamo stimare. Pertanto

$$\hat{p} = \frac{1}{120}\sum_{i=1}^{120} A_i = \frac{1}{120}(\sharp\{1.75 \leq X_i < 1.90\} + \sharp\{X_i \geq 1.90\}) = \frac{1}{120}(36 + 25) = 0.5083333$$

(b) Dato che $0.4*120 = 48$ e $0.6*120 = 72$ entrambi maggiori di 5, possiamo utilizzare un'approssimazione normale; la regione critica è data da

$$R.C. = \left\{(x_1, \ldots, x_n) : |\hat{p} - 0.4|/\sqrt{0.4(1-0.4)/n} \ > \ 1.96\right\}$$

ovvero, si rifiuta $H0$ a livello $\alpha = 0.05$ se $|\hat{p} - 0.4|/\sqrt{0.4(1-0.4)/n} \ > \ 1.96$.

(c) Dato che $0.4*120 = 48$ e $0.6*120 = 72$ entrambi maggiori di 5, possiamo utilizzare un'approssimazione normale. Allora

$$\text{p} - \text{value} = 2(1 - \Phi\Big[\Big|\frac{\hat{p} - 0.4}{\sqrt{\frac{0.4*(1-0.4)}{n}}}\Big|\Big]) \simeq 0.015.$$

Se effettuiamo un test al livello 5%, rifiutiamo H0.

(d) Al punto (b) abbiamo rifiutato H0. Pertanto, se abbiamo preso la decisione sbagliata, abbiamo commesso un errore del I tipo (abbiamo rifiutato H0 quando H0 era in realtà vera).

(e)

$$\beta = \Phi\Big[\big(0.4 - 0.5 + \mathcal{Z}_{0.025}\sqrt{0.4(1-0.4)/120}\big)/\sqrt{0.5(1-0.5)/120}\Big] -$$
$$\Phi\Big[\big(0.4 - 0.5 - \mathcal{Z}_{0.025}\sqrt{0.4(1-0.4)/120}\big)/\sqrt{0.5(1-0.5)/120}\Big] \simeq 0.39$$

Problema 2.

Una coppia di fidanzati vuole comprare casa prima di sposarsi. Devono scegliere se andare a vivere a Lodi o a Pavia, rispettive città natali. Per prendere questa decisione i due fidanzati iniziano a visitare 6 case in centro a Lodi e 8 case in centro a Pavia. I costi al metro quadro delle case visitate sono

Lodi :	2030 2540 2270 2490 2190 2710
Pavia :	3120 3110 3280 3280 2340 2460 2740 3600

1. Osservando i Normal probability plot dei due dataset, mostrati in Figura 1, e sapendo che il p-value del test di Shapiro-Wilk per i prezzi a Lodi e Pavia sono rispettivamente 0.9033 e 0.5404, si può assumere che i dati delle due popolazioni siano normali?

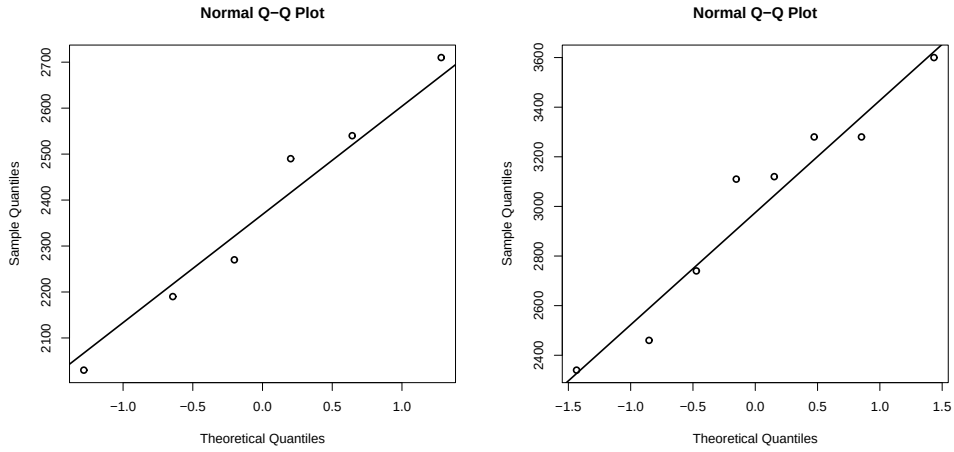

Figura 1: Normal probability plot relativi ai prezzi al metro quadro a Lodi (sinistra) e Pavia (destra).

2. 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.
3. Basandosi sul risultato del test al punto precedente, calcolare un intervallo di confidenza di livello 95% per la differenza delle medie dei prezzi delle case a Lodi e a Pavia.
4. E' possibile affermare ad un livello di significatività del 5% che i prezzi medi delle case nelle due città sono diversi?

Soluzione.

1. I dati nei Normal probability plot sembrano avere un andamento lineare ed il p-value del test di Shapiro-Wilk è molto alto in entrambi i casi. Per questo motivo i dati possono essere considerati normali: $X_L \sim N(\mu_L, \sigma_L^2)$, $X_P \sim N(\mu_P, \sigma_P^2)$.

2. Si vuole effettuare il test $H_0 : \sigma_L = \sigma_P$ vs. $H_1 : \sigma_L \neq \sigma_P$. La regione critica di questo test è

$$R = \left\{ (x_1^L, \ldots, x_{n_L}^L), (x_1^P, \ldots, x_{n_P}^P) : \frac{s_L^2}{s_P^2} > f_{\alpha/2, n_L-1, n_P-1} \text{ or } \frac{s_L^2}{s_P^2} < f_{1-\alpha/2, n_L-1, n_P-1} \right\}.$$

   Dato che varianze campionarie sono $s_L^2 = 63376.67$ e $s_P^2 = 191069.6$, si ottiene $f_0 = s_L^2/s_P^2 = 0.3317$. 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%.

3. 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

$$IC_{0.95} = \left(\bar{x}_L - \bar{x}_P \pm t_{\alpha/2, n_L+n_P-2} s_p \sqrt{\frac{1}{n_L} + \frac{1}{n_P}} \right)$$

   La varianza pooled è pari $s_p^2 = \frac{(n_L-1)s_L^2 + (n_P-1)s_P^2}{n_L+n_P-2} = 137864.2$, $s_p \sqrt{\frac{1}{n_L} + \frac{1}{n_P}} = 200.52$ e $t_{0.025,12} = 2.179$ per cui si ottiene

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

4. 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 la media dei prezzi nelle due città è diversa.

Problema 3.
Il Politecnico di Milano intende presentare ai prossimi campionati universitari, per la specialità del salto in alto femminile, una squadra di 8 studentesse atlete. L'allenatore della squadra decide di monitorare il peso delle atlete, per dimostrare come l'altezza del salto effettuato nella specialità del salto in alto (espresso in centrimetri) dipenda linearmente dal peso dell'atleta (in Kg), ovvero dalla sua massa muscolare. I dati rilevati sono:

salto (cm)	170.34	157.42	171.89	179.35	186.22	183.60	193.40	190.95
peso (Kg)	62.1	57.1	60.3	62.8	65.3	67.0	70.3	68.4

Sotto viene anche riportato l'output di alcune elaborazioni ed analisi di questi dati effettuate con il sofware statistico R.

(a) Si scriva il modello lineare che lega il peso dell'atleta all'altezza del salto effettuato, e si indichino le ipotesi alla base del modello.

(b) Si scriva la retta di regressione stimata.

(c) Si commenti il modello di regressione alla luce dell'output e della diagnostica dei residui in Figura 2.

(d) L'atleta più forte della squadra risulta infortunata al momento del monitoraggio, ma rientrerà appena prima delle gare. L'allenatore vuole comunque avere un'idea della sua possibile prestazione, anche a fronte di una possibile perdita di massa muscolare dovuta all'inattività forzata. Dato che il suo peso attuale è 66 Kg, si fornisca una stima puntuale e un intervallo di previsione al 95% per l'altezza del salto dell'atleta infortunata.

```
> modello=lm(salto  ~ peso)
> summary(modello)

Call:
lm(formula = salto ~ peso)

Residuals:
    Min      1Q  Median      3Q     Max
-3.3792 -3.0451 -0.6212  3.1276  4.0806

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   10.315     19.351   0.533 0.613159
peso           2.631      0.301   8.743 0.000124 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.508 on 6 degrees of freedom
Multiple R-squared: 0.9272,     Adjusted R-squared: 0.9151
F-statistic: 76.43 on 1 and 6 DF,  p-value: 0.0001239

> var(peso)
[1] 19.41125
```

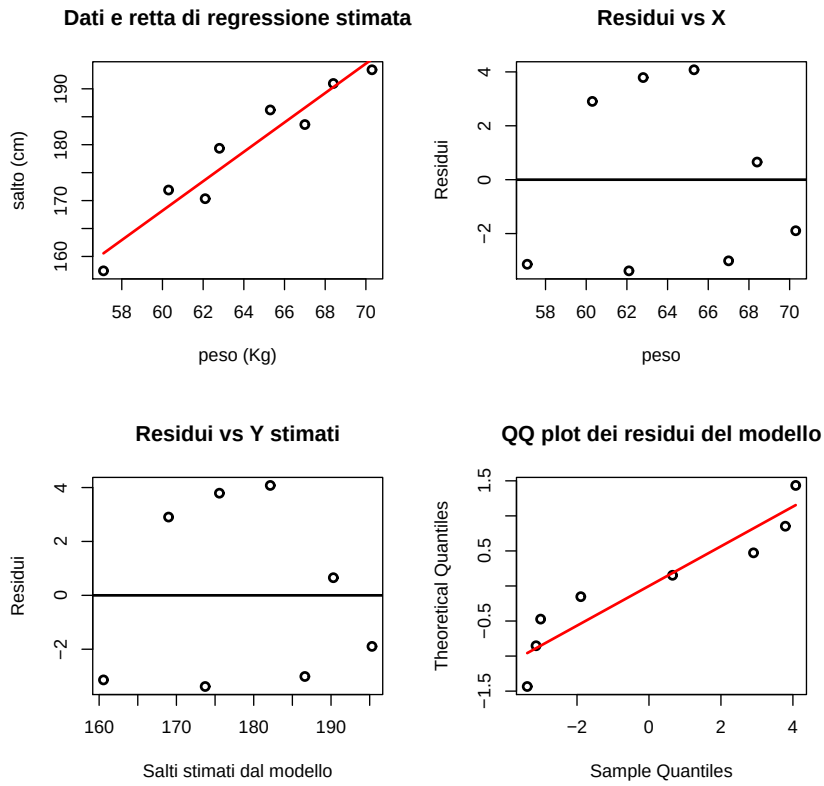

Figura 2: Retta stimata e diagnostica dei residui

Soluzione.

(a) Se indichiamo con Y l'altezza del salto dell'atleta (cm), e con X il suo peso (Kg), il modello lineare ipotizzato è il seguente: $Y = \beta_0 + \beta_1 \cdot X + \epsilon$, dove assumiamo $\epsilon \sim N(0, \sigma^2)$.

(b) $\hat{Y} = 10.3 + 2.63 \cdot X$.

(c) Il modello sembra buono: il test F è molto significativo, e inoltre R^2 è pari a 92.72%, un valore molto alto che dimostra l'ottimo adattamento dei dati al modello ipotizzato. Inoltre i residui non sembrano presentare andamenti anomali (a parte forse una leggera deviazione dall'ipotesi di normalità, comunque non valutabile con un campione così poco numeroso). Sembra tuttavia che l'intercetta del modello lineare non sia significativa.

(d) Avendo fissato $X_0 = 66$, la stima puntuale dell'altezza del salto risulta essere: $\hat{Y}_0 = 10.3 + 2.63 \cdot 66 = 183.88$, mentre l'intervallo di previsione al 95% è dato da:

$$\begin{aligned} IC_{Y_0|X_0}(0.95) &= \left[\hat{Y}_0 \pm t_{0.975,6} \cdot \sqrt{\hat{\sigma}^2 \cdot (1 + 1/8 + (66 - 64.162)^2/135.88)}\right] \\ &= [183.88 \pm 2.45 \cdot 3.76] = [183.88 \pm 9.21]\,. \end{aligned}$$