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

Seconda prova in itinere di Statistica per Ingegneria Energetica
13 febbraio 2015

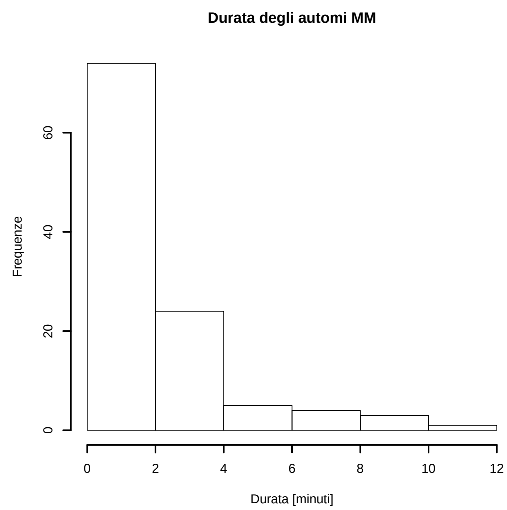
Cognome, Nome e Numero di matricola:

Problema 1. Il Dottor H. vuole inferire sulla durata X di un automa MM. A tal fine registra le durate in minuti di 111 automi MM indipendenti. I dati raccolti hanno media campionaria e varianza campionaria

$$\overline{x}_{111} = 1.96, \qquad s^2_{111} = 4.9902,$$

e, relativamente alle classi di seguito introdotte, risultano così distribuiti

Classi	Frequenze Assolute
$[0-2)$	74
$[2-4)$	24
$[4-6)$	5
$[6-8)$	4
$[8-10)$	3
$[10-12)$	1



(a) Fornire una stima puntuale e una stima intervallare di confidenza 0.95 della durata media $\mu = \mathbb{E}[X]$ di un automa MM.

Il Dottor H. sospetta che la durata X di un automa MM abbia distribuzione esponenziale.

(b) Impostare un opportuno test di adattamento per verificare se X abbia distribuzione esponenziale. Esplicitare le ipotesi statistiche e la regione critica di livello α, utilizzabile con i dati a disposizione.

(c) Calcolare il p-value dei dati raccolti e trarre le dovute conclusioni circa la distribuzione di X.

(d) Fornire una stima puntuale della probabilità che un automa MM duri più di 12 minuti.

Risultati.

(a) Stima puntuale: $\widehat{\mu} = \overline{x}_{111} = 1.96$.

Stima intervallare di confidenza 0.95 (anche se il campione non proviene da una normale, è tuttavia sufficientemente numeroso, essendo $n = 111 > 40$):

$$\widehat{\mu} \pm \frac{s_{111}}{\sqrt{111}}\, t_{0,025}(110) = 1.96 \pm \sqrt{\frac{4.9902}{111}}\, 1.9818 = 1.96 \pm 0.42 = (1.54\,,\,2.38).$$

(b) $H_0 \,:\, X \sim \mathcal{E}, \quad H_1 \,:\, X \nsim \mathcal{E}$.

La regione critica R_α dipende dalla scelta delle classi A_ℓ ed è utilizzabile se

$$n\,\widehat{p}_\ell = 111 \cdot \widehat{\mathbb{P}}(X \in A_\ell) > 5 \quad \text{per ogni } \ell.$$

Le probabilità vanno stimate con la distribuzione esponenziale di media $\widehat{\mu} = 1.96$, ovvero di parametro $\lambda = 1/\widehat{\mu} = 0.510204082$. Per le classi assegnate si ha

A_ℓ	N_ℓ	$\widehat{p}_\ell$	$n\,\widehat{p}_\ell$
$[0\,,\,2)$	74	0.6395522114	70.9902954656
$[2\,,\,4)$	24	0.2305251803	25.5882950125
$[4\,,\,6)$	5	0.0830922915	9.2232443512
$[6\,,\,8)$	4	0.0299504327	3.3244980301
$[8\,,\,10)$	3	0.0107955672	1.1983079631
$[10\,,\,12)$	1	0.0038912383	0.4319274554
$[12\,,\,+\infty)$	0	0.0021930786	0.243431722

per cui conviene accorpare le ultime quattro classi

A_ℓ	N_ℓ	$\widehat{p}_\ell$	$n\,\widehat{p}_\ell$
$[0\,,\,2)$	74	0.6395522114	70.9902954656
$[2\,,\,4)$	24	0.2305251803	25.5882950125
$[4\,,\,6)$	5	0.0830922915	9.2232443512
$[6\,,\,\infty)$	8	0.0468303169	5.1981651706

e utilizzare la corrispondente regione critica $R_\alpha \,:\, Q = \displaystyle\sum_{\ell=1}^{4} \frac{(N_\ell - n\widehat{p}_\ell)^2}{n\widehat{p}_\ell} > \chi^2_\alpha(2)$.

(c) Il p-value dei dati raccolti è la soluzione α di

$$\chi^2_\alpha(2) = Q = \sum_{\ell=1}^{4} \frac{(N_\ell - n\widehat{p}_\ell)^2}{n\widehat{p}_\ell} = 3.67$$

Con le tavole

$$\chi^2_{0.5}(2) = 1.39 < \chi^2_\alpha(2) = 3.67 < \chi^2_{0.1}(2) = 4.61$$

$$0.1 < \alpha < 0.5$$

mentre il valore esatto è $\alpha = 0.16$. Il p-value non è particolarmente alto, ma comunque superiore a 0.1. Pertanto, agli usuali livelli di significatività, non possiamo rifiutare H_0 e concludiamo che la distribuzione di X è un'esponenziale. La conclusione tuttavia è debole.

(d) Coerentemente con la conclusione del test stimiamo $\mathbb{P}(X > 12)$ con

$$\widehat{\mathbb{P}}(X > 12) = \mathrm{e}^{-12/\widehat{\mu}} = 0.0021930786 = 0.22\%.$$

Problema 2. Il Dottor H. vuole confrontare il carico di rottura μ_M degli elmetti M con il carico di rottura μ_S degli elmetti S. In particolare si domanda se si possa ritenere, con forte evidenza statistica, che $\mu_M > \mu_S$. Pertanto il Dottor H. convoca il Barone A. e, dotandolo di uno strumento affetto da errore di misura gaussiano standard (fissata opportunamente l'unità di misura), gli ordina m misure indipendenti di μ_M, con risultati quindi

$$X_1, \dots, X_m \qquad \text{campione casuale} \qquad N(\mu_M, 1),$$

e n misure indipendenti di μ_S, con risultati quindi

$$Y_1, \dots, Y_n \qquad \text{campione casuale} \qquad N(\mu_S, 1).$$

(a) Impostare un opportuno test statistico per poter rispondere alla domanda del Dottor H. Esplicitare ipotesi nulla, ipotesi alternativa e regione critica di livello α.

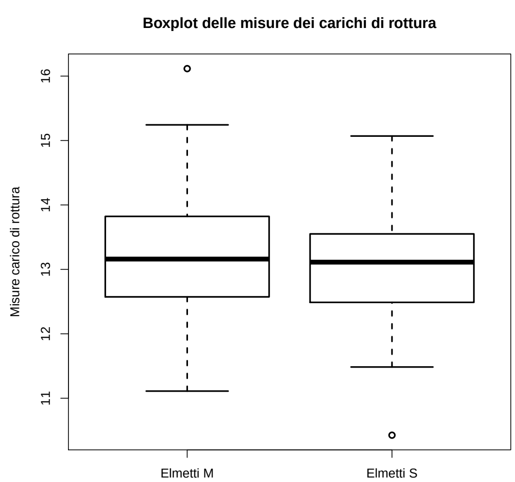
(b) Calcolare la potenza del test introdotto in funzione dei parametri in gioco.

(c) Supponendo $m = n$ e $\alpha = 0.05$, calcolare la minima ampiezza campionaria n capace di rivelare una differenza $\mu_M - \mu_S = 0.5$ con una potenza del 60% almeno.

Alla fine il Dottor H. fa eseguire 49 misure sugli elmetti M e 45 misure sugli elmetti S, ottenendo

$$\overline{x}_{49} = 13.24 \qquad \overline{y}_{45} = 13.05,$$

e i seguenti boxplot



(d) Calcolare il p-value dei dati raccolti.

(e) Trarre le dovute conclusioni agli usuali livelli di significatività.

(f) Se la conclusione fosse sbagliata, avreste commesso un errore del primo o del secondo tipo?

Risultati.

(a) $H_0 \,:\, \mu_M \le \mu_S, \qquad H_1 \,:\, \mu_M > \mu_S, \qquad R_\alpha \,:\, \overline{x}_m > \overline{y}_n + \sqrt{\frac{1}{m} + \frac{1}{n}}\, z_\alpha.$

(b) $\pi = \mathbb{P}\left(\overline{X}_m > \overline{Y}_n + \sqrt{\frac{1}{m} + \frac{1}{n}}\, z_\alpha\right) = \mathbb{P}\left(\frac{\overline{X}_m - \mu_M - \overline{Y}_n + \mu_S}{\sqrt{\frac{1}{m} + \frac{1}{n}}} > z_\alpha - \frac{\mu_M - \mu_S}{\sqrt{\frac{1}{m} + \frac{1}{n}}}\right) = 1 - \Phi\left(z_\alpha - \frac{\mu_M - \mu_S}{\sqrt{\frac{1}{m} + \frac{1}{n}}}\right)$

(c) Supponendo $m = n$, calcolare la minima ampiezza campionaria n capace di rivelare una differenza $\mu_S - \mu_M = 1$ con una potenza del 60% almeno.

$$\pi = 1 - \Phi\left(z_{0.05} - \frac{\mu_M - \mu_S}{\sqrt{\frac{2}{n}}}\right) \ge 0.6$$

$$\Phi\left(z_{0.05} - \frac{\mu_M - \mu_S}{\sqrt{\frac{2}{n}}}\right) \le 0.4$$

$$z_{0.05} - \frac{\mu_M - \mu_S}{\sqrt{\frac{2}{n}}} \le z_{0.6}$$

$$n \ge 2\left(\frac{z_{0.05} - z_{0.6}}{\mu_M - \mu_S}\right)^2 = 2\left(\frac{1.645 + 0.253}{0.5}\right)^2 = 28.8$$

(d) Il p-value dei dati raccolti è la soluzione α di

$$\overline{x}_m = \overline{y}_n + \sqrt{\frac{1}{m} + \frac{1}{n}}\, z_\alpha$$

per cui

$$z_\alpha = \frac{\overline{x}_{49} - \overline{y}_{45}}{\sqrt{\frac{1}{49} + \frac{1}{45}}} = \frac{0.19\sqrt{2205}}{\sqrt{94}} = 0.92$$

$$\alpha = 1 - \Phi(0.92) = 1 - 0.821214 = 0.178786.$$

(e) Nonostante $\overline{x}_{49} > \overline{y}_{45}$, abbiamo un p-value > 0.1 e quindi non possiamo rifiutare l'ipotesi nulla agli usuali livelli di significatività: $\mu_M \le \mu_S$

(f) Secondo tipo.

Problema 3. Il Dottor H. ha bisogno di trovare un buon modello lineare empirico gaussiano che spieghi l'energia Y consumata al minuto da un automa MM con la sua altezza x_1 e la sua larghezza x_2. Pertanto ordina al Barone A. di raccogliere i dati relativi a 21 differenti automi MM, per poi elaborarli con due regressioni lineari multiple: Y su x_1 e x_2 (modello 1) e Y su x_1, x_2 e $x_1 x_2$ (modello 2). Per ciascun modello è allegato lo specchietto riassuntivo della regressione, alcuni grafici dei residui standardizzati, il p-value dei residui standardizzati per il test di normalità di Shapiro-Wilk.

(a) Scrivere il legame fra le variabili Y, x_1 e x_2 ipotizzato dai due modelli lineari empirici gaussiani.

(b) Spiegare quale dei due modelli è il migliore spiegando tutti i pro e contro.

(c) Stimare puntualmente il consumo medio di energia al minuto degli automi MM alti 13 e larghi 8.

(d) Stimare puntualmente la variazione media di energia consumata al minuto passando da automi MM alti 13 e larghi 8 ad automi MM alti 15 e larghi 8.

(e) Il Dottor H. ritiene tuttavia che l'intercetta del modello debba valere 50. Stabilire con un opportuno test di livello 1% se i dati possono confutare tale convinzione. Esplicitare ipotesi statistiche, regione critica e conclusione.

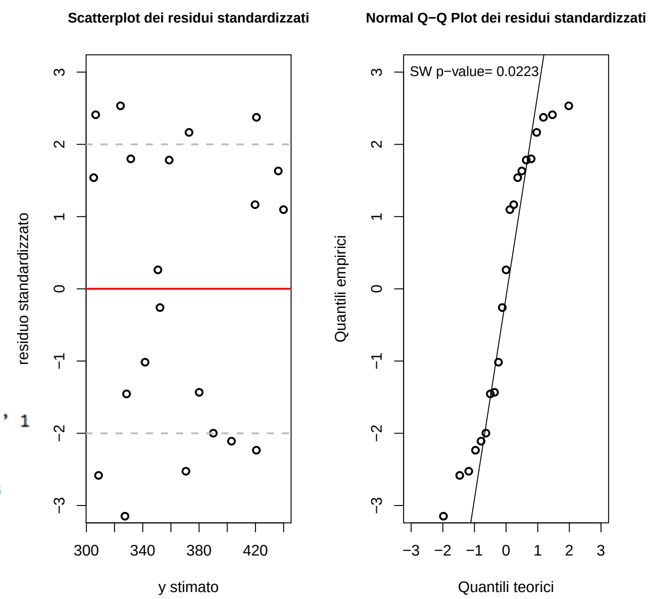
```
Call:
lm(formula = y ~ x1 + x2)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.1481 -1.9980  0.2624  1.7825  2.5336 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  47.0528     3.7324   12.61 2.27e-10 ***
x1           12.6799     0.1504   84.33  < 2e-16 ***
x2           17.8564     0.3994   44.70  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.086 on 18 degrees of freedom
Multiple R-squared:  0.998,	Adjusted R-squared:  0.9978 
F-statistic:  4465 on 2 and 18 DF,  p-value: < 2.2e-16
```

(a) Summary del modello 1



(b) Scatterplot e Q-Q plot dei residui standardizzati del modello 1

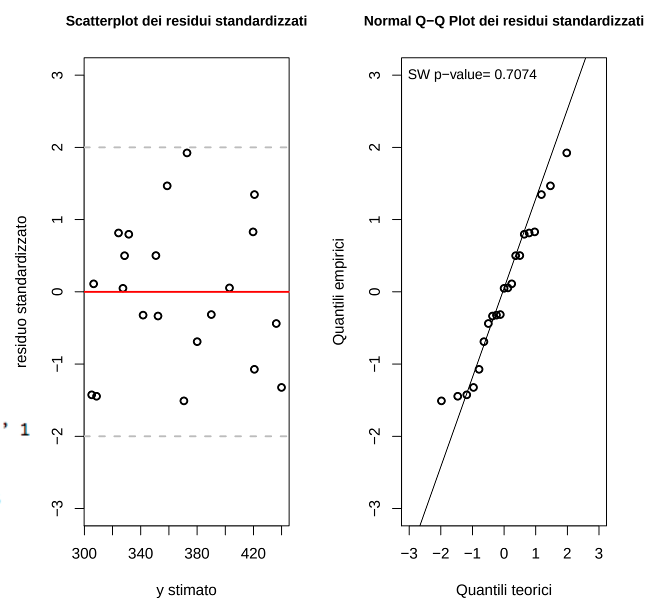
```
Call:
lm(formula = y ~ x1 + x2 + x1:x2)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.69332 -0.78234  0.04564  0.88528  2.24599 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 108.34230   10.25250  10.567 6.85e-09 ***
x1            8.75735    0.64741  13.527 1.58e-10 ***
x2            9.41282    1.40010   6.723 3.57e-06 ***
x1:x2         0.54069    0.08844   6.114 1.15e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.2 on 17 degrees of freedom
Multiple R-squared:  0.9994,	Adjusted R-squared:  0.9993 
F-statistic:  9004 on 3 and 17 DF,  p-value: < 2.2e-16
```

(c) Summary del modello 2



(d) Scatterplot e Q-Q plot dei residui standardizzati del modello 2

Risultati.

(a) Modello 1: $Y = \beta_0 + \beta_1\, x_1 + \beta_2\, x_2 + \epsilon$, dove $\epsilon \sim N(0, \sigma^2)$.

Modello 2: $Y = \beta_0 + \beta_1\, x_1 + \beta_2\, x_2 + \beta_3\, x_1\, x_2 + \epsilon$, dove $\epsilon \sim N(0, \sigma^2)$.

(b) È migliore il modello 2 in quanto:

- è decisamente meglio confermata l'ipotesi gaussiana, dato che
  - entrambi gli scatterplot dei residui standardizzati sono a nuvola senza struttura, ma il modello 1 avrebbe ben 9 outlier su 21 dati, mentre il modello 2 non ha outlier,
  - il normal Q-Q plot dei residui standardizzati è migliore per il modello 2,

    * il p-value di Shapiro-Wilk è molto più alto per il modello 2;

  - è maggiore R^2_{corretto} (0.9978 per il modello 1, 0.9993 per il modello 2), ovvero è minore la stima di σ^2.

Non danno invece indicazioni la significatività globale della regressione (che è la medesima per entrambi i modelli) e le significatività dei singoli regressori (che sono tutte ottime per tutti i regressori per entrambi i modelli).

(c)

$$\begin{aligned}\widehat{\mathbb{E}}[Y|x_1 = 13, x_2 = 8] &= \widehat{\beta}_0 + \widehat{\beta}_1\, x_1 + \widehat{\beta}_2\, x_2 + \widehat{\beta}_3\, x_1\, x_2 \\ &= 108.34230 + 8.75735 \cdot 13 + 9.41282 \cdot 8 + 0.54069 \cdot 13 \cdot 8 = 353.722.\end{aligned}$$

(d)

$$\begin{aligned}\widehat{\mathbb{E}}[Y|x_1 = 15, x_2 = 8] - \widehat{\mathbb{E}}[Y|x_1 = 13, x_2 = 8] &= \widehat{\beta}_1 \cdot (15 - 13) + \widehat{\beta}_3 \cdot (15 - 13) \cdot 8 \\ &= 8.75735 \cdot 2 + 0.54069 \cdot 2 \cdot 8 = 25.80.\end{aligned}$$

(e) $H_0 \,:\, \beta_0 = 50, \quad H_1 \,:\, \beta_0 \neq 50, \quad R_\alpha \,:\, |\widehat{\beta}_0 - 50| > \text{se}(\widehat{\beta}_0)\, t_{\alpha/2}(n-4).$

Per i dati raccolti

$$|\widehat{\beta}_0 - 50| = 58.3423 > \text{se}(\widehat{\beta}_0)\, t_{0.005}(17) = 10.25250 \cdot 2.898 = 29.7117$$

quindi ad un livello dell'1% i dati consentono di rifiutare l'ipotesi nulla. Pertanto $\beta_0 \neq 50$.