

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

**APPELLO DI STATISTICA PER INGEGNERIA ENERGETICA
13 febbraio 2015**

©I diritti d'autore sono riservati. Ogni sfruttamento commerciale non autorizzato sarà perseguito.

Cognome, Nome e Numero di matricola:

Problema 1. Aureliano è uno studente fuori sede che vive a Macondo, uno sperduto paesino immerso nelle foreste pluviali a nord di Milano. Ogni giorno, per andare in università, Aureliano prende l'unico treno che parte dalla stazione di Macondo alle 7:00 e per tornare a casa dopo le lezioni l'unico treno che torna da Milano alle 17:00.

Il tempo di percorrenza delle tratte Macondo-Milano e Milano-Macondo (in ore) sono modellizzabili con la stessa variabile aleatoria continua T di densità:

$$f_T(t) = \frac{3}{t^4} I_{(1,\infty)}(t) = \begin{cases} 0 & t \leq 1 \\ \frac{3}{t^4} & t > 1 \end{cases}$$

Aureliano, che purtroppo è un po' debole in statistica, vi chiede di aiutarlo a rispondere alle domande seguenti.

- (a) Disegnare un grafico qualitativo di $f_T(t)$, e calcolare media e varianza del tempo impiegato da Aureliano per fare un viaggio in treno.
- (b) Calcolare la funzione di ripartizione $F_T(t)$ e tracciarne un grafico qualitativo.
- (c) Domani Aureliano deve sostenere un esame ed è quindi molto importante che riesca ad arrivare a Milano entro le 9:00. Calcolare la probabilità che Aureliano arrivi a Milano in tempo.

A causa dei continui ritardi, le ferrovie macondiane da tempo riconoscono ai pendolari un bonus. In particolare, se in un mese (che si suppone essere di 30 giorni) il tempo totale di percorrenza dei 60 treni da e per Milano supera le 100 ore, il mese successivo tutti gli abbonamenti mensili vengono venduti a metà prezzo. Si suppone che i tempi di percorrenza di viaggi diversi siano tra loro indipendenti.

- (d) Calcolare la probabilità, eventualmente approssimata, che Aureliano in un mese qualsiasi abbia diritto al bonus.
- (e) Sapendo che Aureliano compra l'abbonamento mensile 11 mesi all'anno, calcolare la probabilità che in un anno abbia diritto al bonus almeno una volta.

Risultati.

(a) Valore atteso: $E[T] = \int_1^\infty t f_T(t) dt = \int_1^\infty t \frac{3}{t^4} dt = [-\frac{3}{2}t^{-2}]_1^\infty = \frac{3}{2}$.

Varianza: $E[T^2] = \int_1^\infty t^2 \frac{3}{t^4} dt = [-3t^{-1}]_1^\infty = 3 \rightarrow Var(T) = 3 - \left(\frac{3}{2}\right)^2 = 0.75$.

(b)

$$F_T(t) = \int_{-\infty}^t f_T(u) du = \begin{cases} 0 & t \leq 1 \\ [-\frac{3}{2}u^{-2}]_1^t = 1 - t^{-2} & t > 1 \end{cases}$$

(c) $P[T \leq 2] = F_T(2) = 1 - 2^{-2} = 0.75$.

(d) Sia X la v.a. che descrive il tempo totale di percorrenza di 60 viaggi: $X = \sum_{i=1}^{60} T_i$, con $T_i \sim F_T$ i tempi di percorrenza dei singoli viaggi: $E[X] = 60E[T_i] = 90$, $Var(X) = 60Var(T_i) = 45$. Essendo il campione numeroso, per il TCL si ha: $X \simeq N(90, 45)$, quindi la probabilità di ottenere il bonus è:

$$P[X > 100] = P\left[\frac{X - 90}{\sqrt{45}} > \frac{100 - 90}{\sqrt{45}}\right] \simeq 1 - \Phi(1.49) = 0.0681.$$

(e) Vista l'indipendenza tra i tempi di percorrenza dei singoli viaggi, anche i tempi di percorrenza mensili totali saranno tra loro indipendenti. Quindi la probabilità cercata è $1 - P[X \leq 100]^{11} = 0.5397$.

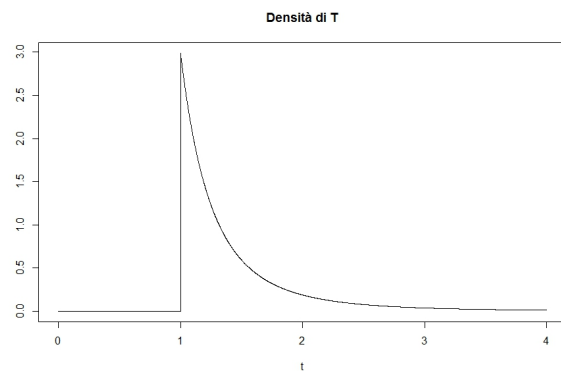


Figura 1: Grafico della funzione densità $f_T(t)$

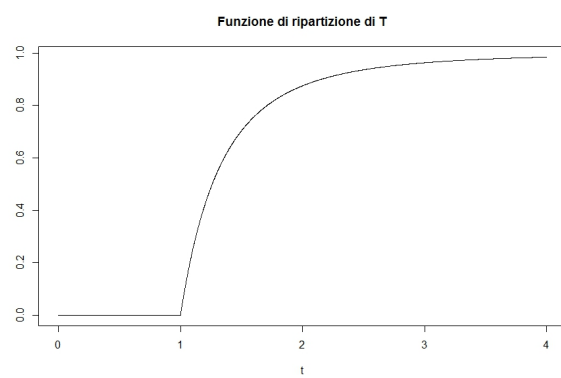


Figura 2: Grafico della funzione di ripartizione $F_T(t)$

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

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

e n misure indipendenti di μ_S , con risultati quindi

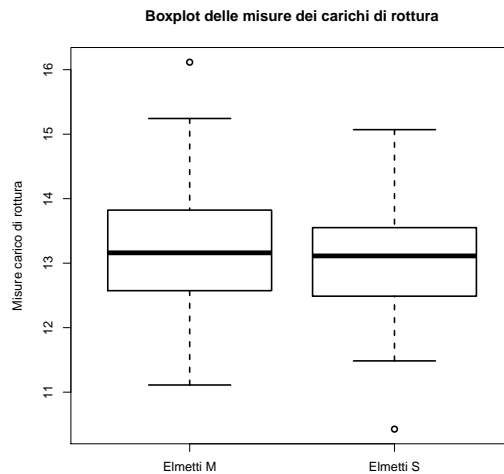
$$Y_1, \dots, Y_n \quad \text{campione casuale} \quad N(\mu_S, 1).$$

- Impostare un opportuno test statistico per poter rispondere alla domanda del Dottor H. Esplicitare ipotesi nulla, ipotesi alternativa e regione critica di livello α .
- Calcolare la potenza del test introdotto in funzione dei parametri in gioco.
- 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

$$\bar{x}_{49} = 13.24 \quad \bar{y}_{45} = 13.05,$$

e i seguenti boxplot



- Calcolare il p-value dei dati raccolti.
- Trarre le dovute conclusioni agli usuali livelli di significatività.
- Se la conclusione fosse sbagliata, avreste commesso un errore del primo o del secondo tipo?

Risultati.

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

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

(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.

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

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

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

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

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

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

per cui

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

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

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

(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.

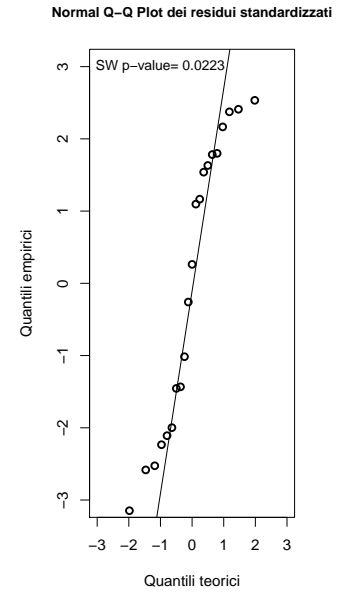
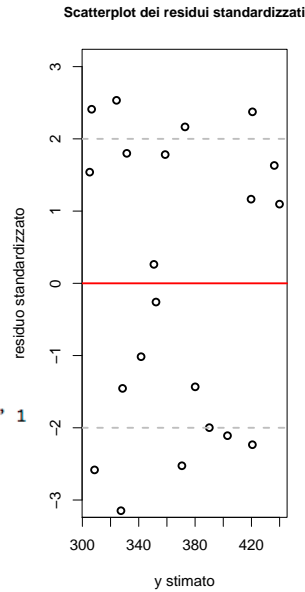
```
Call:
lm(formula = y ~ x1 + x2)

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

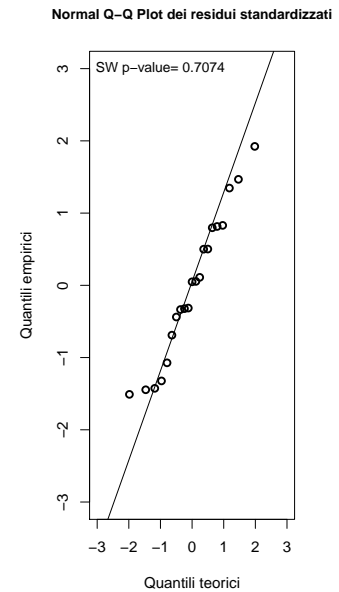
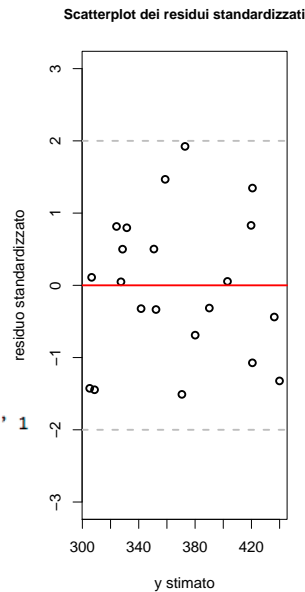
```
Call:
lm(formula = y ~ x1 + x2 + x1:x2)

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_{corretto}^2 (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 $\beta_0 \neq 50$.