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

I Appello di Statistica per Ingegneria Fisica
3 Luglio 2018

Cognome, Nome e Numero di matricola:

Problema 1. Ogni lezione di Statistica del Professor T. inizia a un'ora X, che è una variabile aleatoria gaussiana centrata intorno alle 8:00 ($= 8$). Inoltre, il 90% delle sue lezioni inizia dopo le 7:30 ($= 7.5$). Infine, le ore d'inizio di lezioni diverse sono tutte indipendenti tra loro.

(a) Determinate la varianza di X. *[Se non ci riuscite, d'ora in poi assumete* $\mathrm{Var}(X) = 0.2$.*]*

Quest'anno Aldo deve frequentare 60 lezioni di Statistica. Poiché è un ragazzo puntuale e ci tiene molto ad arrivare prima del Prof. T., ha deciso che tutte le mattine si presenterà in aula esattamente alle 7:30.

(b) In questo modo, qual è la probabilità che non perda mai l'inizio di una lezione?

(c) E qual è invece la probabilità che ne perda al massimo 10?

Anche l'ora a cui finisce una qualsiasi lezione del Prof. T. è una variabile aleatoria Y gaussiana e con la stessa varianza di X. Tuttavia, Y è centrata intorno alle ore 9:00 ($= 9$) anziché alle 8:00.

(d) Sia D la durata di una lezione del Prof. T.. Calcolare media e varianza di D.

(e) Qual è la probabilità che una lezione del Prof. T. duri più di un'ora?

Risultati.

(a) Ponendo $\mu = \mathbb{E}[X] = 8$ e $\sigma^2 = \mathrm{Var}(X)$, sappiamo che

$$0.90 \equiv \mathbb{P}(X > 7.5) = \mathbb{P}\left(\underbrace{\frac{X-\mu}{\sigma}}_{\sim N(0,1)} > \frac{7.5-8}{\sigma}\right) = 1 - \Phi\left(\frac{7.5-8}{\sigma}\right) = \Phi\left(\frac{0.5}{\sigma}\right)$$

$$\Rightarrow \quad \frac{0.5}{\sigma} = z_{0.90} = 1.28 \quad \Rightarrow \quad \sigma = \frac{0.5}{1.28} = 0.390625$$

$$\Rightarrow \quad \mathrm{Var}(X) = 0.390625^2 = 0.152588\,.$$

(b) Se chiamiamo $S =$ numero di lezioni perse da Aldo, allora $S \sim B(60, (1-0.90)) = B(60, 0.10)$, e la probabilità cercata è

$$\mathbb{P}(S = 0) = \binom{60}{0}(0.10)^0(1-0.10)^{60-0} = 0.90^{60} = 0.1797\%\,.$$

(c) Ora vogliamo calcolare

$$\mathbb{P}(S \le 10) = \mathbb{P}(S \le 10.5) = \mathbb{P}\left(\underbrace{\frac{S - \mathbb{E}[S]}{\sqrt{\mathrm{Var}(S)}}}_{\substack{\approx N(0,1) \\ \text{per il TLC}}} \le \frac{10.5 - 60\cdot 0.10}{\sqrt{60\cdot 0.10\cdot(1-0.10)}}\right)$$

$$\simeq \Phi\left(\frac{10.5 - 60\cdot 0.10}{\sqrt{60\cdot 0.10\cdot(1-0.10)}}\right) = \Phi(1.936) \simeq 0.9735 = 97.35\%\,.$$

Senza correzione di continuità:

$$\mathbb{P}(S \le 10) = \mathbb{P}\left(\frac{S - \mathbb{E}[S]}{\sqrt{S}} \le \frac{10 - 60\cdot 0.10}{\sqrt{60\cdot 0.10\cdot(1-0.10)}}\right)$$

$$\simeq \Phi\left(\frac{10 - 60\cdot 0.10}{\sqrt{60\cdot 0.10\cdot(1-0.10)}}\right) = \Phi(1.721) = 0.9573 = 95.73\%\,.$$

(d) Abbiamo

$$D = Y - X$$

e quindi

$$\mathbb{E}[D] = \mathbb{E}[Y] - \mathbb{E}[X] = 9 - 8 = 1$$
$$\mathrm{Var}(D) \overset{\text{indip.}}{=} \mathrm{Var}(Y) + \mathrm{Var}(X) \simeq 0.152588 + 0.152588 = 0.305176$$

(e) Poiché X e Y sono gaussiane e indipendenti, abbiamo

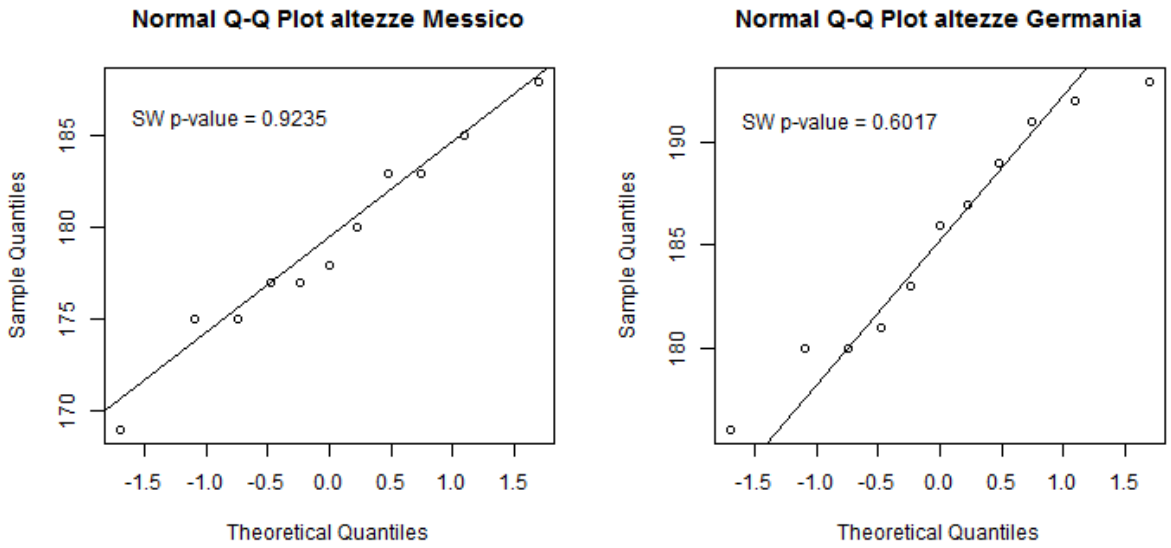
$$D \sim N(\mathbb{E}[D], \mathrm{Var}(D)) = N(1, 0.305176).$$

Pertanto,

$$\mathbb{P}(D > 1) = \mathbb{P}(D > \mathbb{E}[D]) = \frac{1}{2} \qquad \text{perché } N(\mathbb{E}[D], \mathrm{Var}(D)) \text{ è simmetrica intorno a } \mathbb{E}[D].$$

Problema 2. Durante la partita di calcio Germania-Messico di tre settimane fa, il commentatore ha affermato che "la media campionaria delle altezze dei giocatori messicani è nettamente minore dell'analoga media per i giocatori tedeschi". Incuriositi da questa affermazione, ci siamo chiesti se tale differenza costituisca una prova del fatto che l'altezza media di tutti i cittadini messicani sia minore di quella dei loro omologhi tedeschi, oppure no. Abbiamo pertanto cercato su Internet le altezze degli 11 titolari messicani e tedeschi che hanno giocato la partita, e le abbiamo poi elaborate con R ottenendo i valori delle medie e delle varianze campionarie riassunti nella tabella seguente. Sotto, sono riportati anche il normal Q-Q plot col p-value del test di Shapiro-Wilk per ciascuno dei due campioni.

	media campionaria (in cm)	varianza campionaria (in cm^2)
titolari messicani	179.09	29.091
titolari tedeschi	185.27	32.018



(a) Con un test al livello di significatività del 10%, decidete se le varianze delle altezze dei cittadini messicani e tedeschi possano considerarsi uguali oppure no. Scrivete esplicitamente le ipotesi statistiche, la statistica test e la regione di rifiuto che portano alla vostra conclusione. Tale conclusione è debole o forte?

(b) Quali sono le ipotesi alla base del test del punto (a)? Sono verificate?

(c) Con un test al livello di significatività del 2.5%, decidete se i dati dimostrano che l'altezza media dei cittadini messicani è effettivamente minore di quella dei cittadini tedeschi. Anche qui, scrivete le ipotesi statistiche, la statistica test e la regione di rifiuto. La vostra conclusione è debole o forte?

(d) Quali sono le ipotesi alla base del test del punto (c)? Sono verificate?

(e) Calcolate il p-value del test del punto (c) o almeno determinate un intervallo in cui esso è compreso.

Risultati.

(a) Impostiamo un test bilatero per il rapporto delle varianze di due popolazioni gaussiane. Le ipotesi statistiche da mettere a confronto sono

$$H_0 : \sigma_X^2 = \sigma_Y^2 \qquad \text{vs.} \qquad H_1 : \sigma_X^2 \neq \sigma_Y^2 \,,$$

dove σ_X^2 e σ_Y^2 sono le vere varianze (naturalmente incognite) della popolazione messicana e tedesca, rispettivamente. La regola di un test al livello α per tali ipotesi è

$$\text{"rifiuto } H_0 \text{ se } F_0 < f_{\frac{\alpha}{2}}(m-1, n-1) \text{ oppure } F_0 > f_{1-\frac{\alpha}{2}}(m-1, n-1)\text{"} \,,$$

dove F_0 è la statistica test

$$F_0 := \frac{S_X^2}{S_Y^2}\,.$$

Coi nostri dati

$$f_0 = \frac{29.091}{32.018} = 0.90858\,,$$

e con $\alpha = 10\%$

$$f_{\frac{\alpha}{2}}(m-1, n-1) = f_{0.05}(10,10) = \frac{1}{f_{1-0.05}(10,10)} = \frac{1}{2.978} = 0.3358$$
$$f_{1-\frac{\alpha}{2}}(m-1, n-1) = f_{0.95}(10,10) = 2.978\,.$$

Poiché

$$0.3358 < f_0 = 0.90858 < 2.978\,,$$

non possiamo rifiutare l'ipotesi nulla che le due varianze siano uguali. Questa è una conclusione *debole*.

(b) Le ipotesi alla base del test precedente sono che le altezze $X_1, \ldots, X_{11}$ degli 11 giocatori tedeschi e le altezze $Y_1, \ldots, Y_{11}$ degli altrettanti giocatori messicani costituiscano due campioni aleatori indipendenti e *normali*. L'indipendenza è ovvia. La normalità, invece, è verificata per via del fatto che in entrambi i normal Q-Q plot i punti sono ben allineati lungo la normal Q-Q line, e il p-value del test di Shapiro-Wilk è elevato ($\gg 5 - 10\%$) per entrambi i campioni.

(c) Ora facciamo un test unilatero per la differenza delle medie μ_X e μ_Y delle popolazioni da cui provengono i due campioni:

$$H_0 : \mu_X = \mu_Y \ (\text{o } \mu_X \geq \mu_Y) \qquad \text{vs.} \qquad H_1 : \mu_X < \mu_Y\,.$$

Abbiamo messo $\mu_X < \mu_Y$ nell'ipotesi alternativa perché questo è ciò che vorremmo dimostrare. Per due campioni così poco numerosi e a varianze incognite, l'unico test che sappiamo fare è il T-test. La sua regola di rifiuto al livello α è

$$\text{“rifiuto } H_0 \text{ se } T_0 < -t_{1-\alpha}(m+n-2)\text{”}\,,$$

dove T_0 è la statistica test

$$T_0 := \frac{\bar{X}_m - \bar{Y}_n}{\sqrt{S_P^2\left(\frac{1}{m} + \frac{1}{n}\right)}} \qquad \text{con} \qquad S_P^2 := \frac{(m-1)S_X^2 + (n-1)S_Y^2}{m+n-2}\,.$$

Abbiamo trovato

$$s_P^2 = \frac{(11-1)\cdot 29.091 + (11-1)\cdot 32.018}{11+11-2} = 30.5545$$
$$t_0 = \frac{179.09 - 185.27}{\sqrt{30.5545 \cdot \left(\frac{1}{11} + \frac{1}{11}\right)}} = -2.6220$$

e con $\alpha = 2.5\%$

$$t_{1-\alpha}(m+n-2) = t_{0.975}(20) = 2.0860\,.$$

Poiché $-2.6220 < -2.0860$, dobbiamo rifiutare H_0 e concludere che al 2.5% di significatività c'è evidenza che l'altezza media dei cittadini messicani è minore di quella dei cittadini tedeschi.

(d) Il T-test del punto precedente si può fare solo se vale l'ipotesi che i due campioni siano indipendenti, normali e abbiano *varianze uguali*. Abbiamo già visto al punto (b) che le ipotesi di indipendenza e normalità sono soddisfatte. Per quanto riguarda la condizione $\sigma_X^2 = \sigma_Y^2$, anch'essa è già stata verificata al punto (a).

(e) Il p-value del test del punto (c) è il valore di α che soddisfa l'uguaglianza

$$t_0 \equiv -t_{1-\alpha}(m+n-2) \quad \Leftrightarrow \quad -2.6220 = -t_{1-\alpha}(20) \quad \Leftrightarrow \quad 2.6220 = t_{1-\alpha}(20)\,.$$

Dal momento che

$$t_{0.99}(20) = 2.5280 < 2.6220 < 2.8453 = t_{0.995}(20)\,,$$

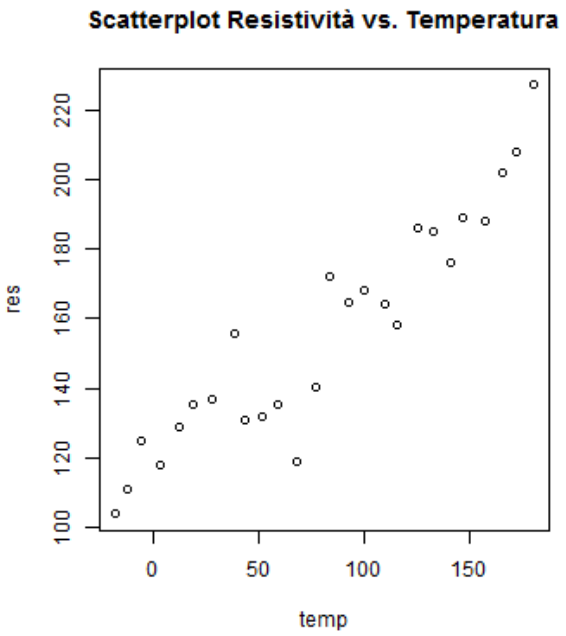
abbiamo

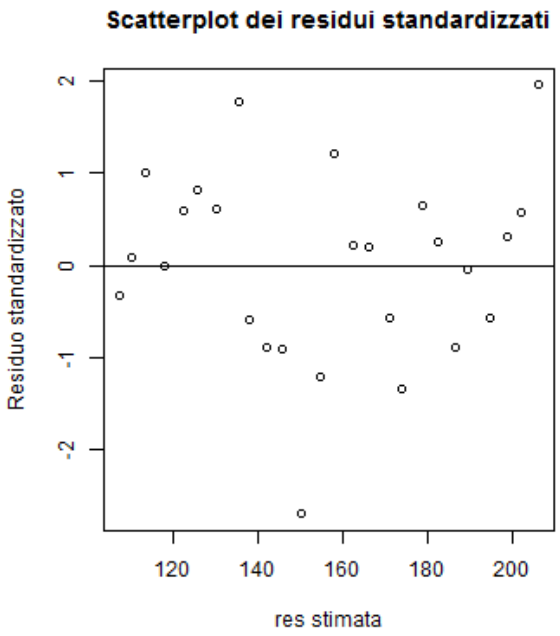
$$0.99 < 1 - p\text{-value} < 0.995 \quad \Leftrightarrow \quad 0.5\% < p\text{-value} < 1\%\,.$$

Problema 3. Nei laboratori della ACME Ltd., i tecnici stanno studiando le proprietà elettriche di una nuova lega metallica destinata alla saldatura dei circuiti elettronici. In particolare, si vuole analizzare la dipendenza della resistività delle lega (variabile $y =$ `res`, espressa in $\mu\Omega \cdot$ mm) in funzione della temperatura (variabile $x =$ `temp`, espressa in °C). È noto infatti che, in un intervallo di temperature sufficientemente ridotto, tale dipendenza è di tipo lineare:

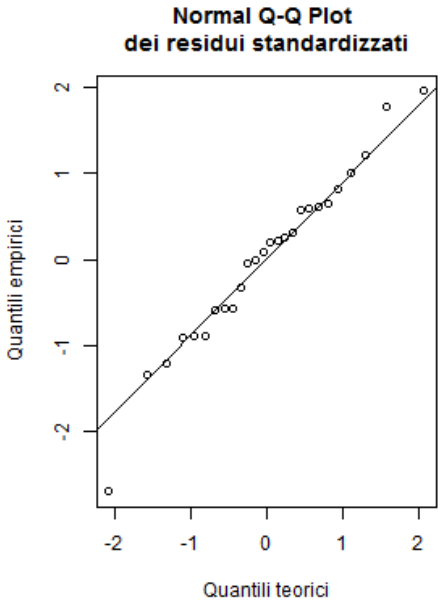
$$y = \beta_0 + \beta_1 x\,.$$

I tecnici hanno dunque misurato la resitività della lega a temperature diverse, ottenendo lo scatterplot dei dati (x_i, y_i) riportato qua sotto. È mostrato anche l'output di R per il modello di regressione lineare considerato. Purtroppo, però, un tecnico ha inavvertitamente fatto cadere alcune gocce di solvente sulla stampa contenente il summary della regressione, rendendo così illeggibili alcune sue parti.






```
> summary(lm(res~temp))

Call:
lm(formula = res ~ temp)

Residuals:
    Min      1Q  [illegible]     3Q     Max
-31.367  -6.833              6.940  21.476

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 116.39715    3.87778  [illegible] < 2e-16 ***
temp          0.49736    0.03868   12.86  [illegible]
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11.9 on 24 degrees of freedom
Multiple R-squared:  [illegible]     Adjusted R-squared:  [illegible]
F-statistic: 165.3 on 1 and 24 DF,  p-value: 2.956e-12

> shapiro.test(rstandard(lm(res~temp)))

        Shapiro-Wilk normality test

data:  rstandard(lm(res ~ temp))
W = 0.97525, p-value = 0.7608
```



(a) Scrivere la relazione tra le variabili y_i e x_i ipotizzata dal modello di regressione empirico gaussiano.

(b) Le ipotesi alla base del modello nel punto (a) vi sembrano soddisfatte? Giustificate adeguatamente la risposta.

(c) Dall'output di R siamo riusciti a ricavare che la variabilità dei dati in uscita era $\sum_i (y_i - \bar{y})^2 = 26\,811.94$. Quale percentuale di questa variabilità è spiegata dal modello lineare?

(d) Fornite una stima puntuale e una stima intervallare al livello di confidenza del 95% per il valor medio della resistività della lega sottoposta a una temperatura di 0°C.

La nuova lega metallica è stata progettata con lo scopo di ridurre la dipendenza della resistività delle saldature dall'aumentare della temperatura nei circuiti. Con le leghe utilizzate finora, infatti, ogni aumento di 1°C della temperatura provocava una variazione media della resistività pari ad almeno $0.60\,\mu\Omega \cdot \text{mm}$.

(e) Con un opportuno test al livello di significatività del 5%, stabilite se, quando la temperatura aumenta di 1°C, la variazione media della resistività della nuova lega è sensibilmente minore di quella delle altre leghe utilizzate finora.

Risultati.

(a) La relazione è

$$Y_i = \beta_0 + \beta_1 x_i + E_i \quad \text{con} \quad E_1, \dots, E_{26} \text{ i.i.d., } E_i \sim N(0, \sigma^2)\ \forall i\,.$$

Notare che i dati sono 26, come si può vedere dai gradi di libertà del test di Fisher nell'output di R o direttamente contando i punti nello scatterplot.

(b) Per verificare se l'ipotesi del punto (a) è soddisfatta, bisogna testare la gaussianità e l'omoschedasticità dei residui standardizzati. Ora, la forma a nuvola del loro scatterplot dimostra l'omoschedasticità. D'altra parte, anche la gaussianità è verificata, come si vede dal fatto che nel normal Q-Q plot i residui si dispongono lungo una retta, e il p-value del test di Shapiro-Wilk è elevato (p-value $= 0.7608$).

(c) La percentuale della variabilità dei dati in uscita spiegata dal modello lineare è il coefficiente di determinazione

$$r^2 = \frac{s_{xy}^2}{s_{xx} s_{yy}} = \hat{\beta}_1^2 \, \frac{s_{xx}}{s_{yy}} = (0.49736)^2 \, \frac{s_{xx}}{26\,811.94}\,,$$

dove abbiamo usato il fatto che $\hat{\beta}_1 = s_{xy}/s_{xx}$. Ci resta da ricavare s_{xx}:

$$\text{se}(\hat{\beta}_1) = \frac{\hat{\sigma}}{\sqrt{s_{xx}}} \quad \Rightarrow \quad s_{xx} = \left(\frac{\hat{\sigma}}{\text{se}(\hat{\beta}_1)} \right)^2 = \left(\frac{11.9}{0.03868} \right)^2 = 94\,650.08\,.$$

La percentuale richiesta è dunque

$$r^2 = (0.49736)^2 \, \frac{94\,650.08}{26\,811.94} = 0.87324 = 87.324\%\,.$$

(d) La resistività media della lega a 0°C è

$$\mathbb{E}\,[Y(0)] = \mathbb{E}\,[\beta_0 + \beta_1 \cdot 0 + E] = \mathbb{E}\,[\beta_0 + E] = \beta_0 + \mathbb{E}\,[E] = \beta_0\,.$$

Pertanto, si richiede di dare una stima puntuale e una intervallare del parametro β_0. La stima puntale è $\hat{\beta}_0 = 116.39715$. La stima intervallare al livello $\gamma = 0.95$ invece è data da

$$\begin{aligned} \beta_0 \in \left(\hat{\beta}_0 \pm t_{\frac{1+\gamma}{2}} (n-k-1)\, \text{se}(\hat{\beta}_0) \right) &= \left(\hat{\beta}_0 \pm t_{0.975}(24)\, \text{se}(\hat{\beta}_0) \right) = (116.39715 \pm 2.0639 \cdot 3.87778) \\ &= (108.394\,,\ 124.401)\,. \end{aligned}$$

(e) La variazione media della resistività in corrispondenza dell'aumento di 1°C della temperatura è

$$\begin{aligned} \mathbb{E}\,[Y(x+1) - Y(x)] &= \mathbb{E}\,[\beta_0 + \beta_1(x+1) + E' - (\beta_0 + \beta_1 x + E)] = \mathbb{E}\,[\beta_1 + E' - E] \\ &= \beta_1 + \mathbb{E}\,[E'] - \mathbb{E}\,[E] = \beta_1\,. \end{aligned}$$

Vogliamo verificare se tale media è sensibilmente minore di 0.60 $\Rightarrow$ dobbiamo fare un test per le ipotesi

$$H_0 : \beta_1 = 0.60 \qquad \text{vs.} \qquad H_1 : \beta_1 < 0.60\,.$$

La regola di rifiuto al livello α è

$$\text{“rifiuto } H_0 \text{ se } \quad \frac{\hat{\beta}_1 - 0.60}{\text{se}(\hat{\beta}_1)} < -t_{1-\alpha}(n-k-1)\text{”}\,.$$

Abbiamo

$$\frac{\hat{\beta}_1 - 0.60}{\text{se}(\hat{\beta}_1)} = \frac{0.49736 - 0.60}{0.03868} = -2.6536$$

e, con $\alpha = 0.05$, $n = 26$ dati e $k = 1$ predittori,

$$t_{1-\alpha}(n-k-1) = t_{0.95}(24) = 1.7109\,.$$

Dal momento che $-2.6536 < -1.7109$, rigettiamo H_0 e al 5% di significatività concludiamo che c'è evidenza che $\beta_1 < 0.60$.