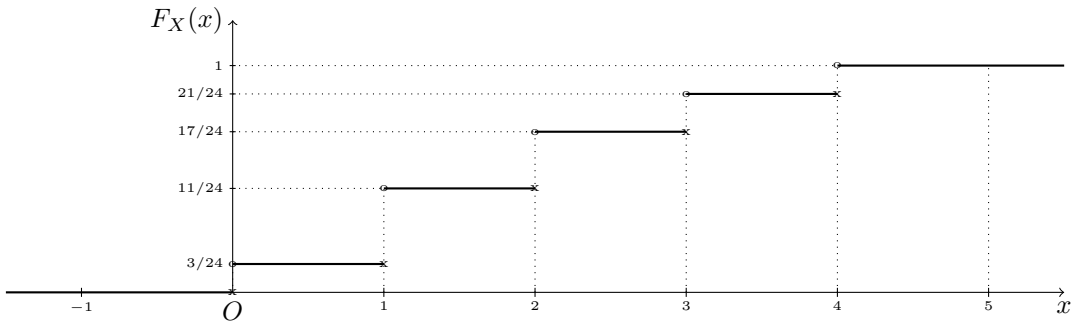
Politecnico di Milano - Scuola di Ingegneria Industriale e dell'Informazione

II appello di Statistica per Ingegneria Fisica
29 Luglio 2019

Cognome, Nome e Numero di matricola:

Problema 1. È una verità incontestabile che, per prepare un buon pesto alla genovese, bisogna usare del basilico raccolto ancora molto piccolo. Lo sa bene il signor Baciccia, che coltiva una serra a Genova Prà. Egli, infatti, raccoglie il suo basilico quando il numero di foglie su ciascuna piantina è una variabile aleatoria X con la funzione di ripartizione disegnata nel grafico seguente:



Naturalmente, mentre crescono, le piantine del signor Baciccia non si influenzano tra loro. Si può dunque assumere che le quantità di foglie su piantine diverse siano tutte indipendenti tra loro.

(a) Qual è il numero massimo di foglie che è possibile trovare su una piantina di basilico del signor Baciccia? E qual è invece il numero minimo?

(b) Determinate la distribuzione di massa di probabilità della variabile aleatoria X, specificando l'insieme su cui è definita tale distribuzione.

(c) Calcolate media e varianza di X.

(d) Calcolate la probabilità che una piantina di basilico del signor Baciccia abbia 3 o più foglie.

(e) Calcolate la probabilità che in un mazzetto di 10 piantine di basilico del signor Baciccia ce ne siano almeno 2 con 3 o più foglie.

(f) Per preparare abbastanza pesto da condirci un piatto di trofie, bisogna usare almeno 80 foglie di basilico. Calcolate la probabilità di riuscirci usando 45 piantine del signor Baciccia.

Risultati.

(a) Dal grafico della funzione di ripartizione si vede che X è una v.a. discreta (ovviamente!) che può prendere con probabilità diversa da zero solo i valori 0, 1, 2, 3 e 4. Di conseguenza, una piantina di basilico del signor Baciccia può avere al massimo 4 foglie e come minimo 0.

(b) Per quanto già visto al punto precedente, l'insieme su cui è definita la densità di probabilità (discreta) di X è $S = \{0, 1, 2, 3, 4\}$. I valori di tale densità sono poi le ampiezze dei 'salti' della funzione di ripartizione, e cioè

$$p(0) = \frac{3}{24} - 0 = \frac{1}{8} \qquad p(1) = \frac{11}{24} - \frac{3}{24} = \frac{1}{3} \qquad p(2) = \frac{17}{24} - \frac{11}{24} = \frac{1}{4}$$
$$p(3) = \frac{21}{24} - \frac{17}{24} = \frac{1}{6} \qquad p(4) = 1 - \frac{21}{24} = \frac{1}{8}$$

(c) Abbiamo

$$\mathbb{E}[X] = \sum_{x\in S} xp(x) = 0\cdot\frac{1}{8} + 1\cdot\frac{1}{3} + 2\cdot\frac{1}{4} + 3\cdot\frac{1}{6} + 4\cdot\frac{1}{8} = \frac{11}{6}$$

$$\mathbb{E}\left[X^2\right] = \sum_{x\in S} x^2p(x) = 0^2\cdot\frac{1}{8} + 1^2\cdot\frac{1}{3} + 2^2\cdot\frac{1}{4} + 3^2\cdot\frac{1}{6} + 4^2\cdot\frac{1}{8} = \frac{29}{6}$$

$$\mathrm{Var}(X) = \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2 = \frac{29}{6} - \left(\frac{11}{6}\right)^2 = \frac{53}{36}$$

(d) La probabilità richiesta è

$$\mathbb{P}(X \geq 3) = \sum_{\substack{x\in S\\ x\geq 3}} p(x) = p(3) + p(4) = \frac{1}{6} + \frac{1}{8} = \frac{7}{24}$$

o, in altro modo,

$$\mathbb{P}(X \geq 3) = 1 - \mathbb{P}(X < 3) = 1 - \mathbb{P}(X \leq 2) = 1 - F(2) = 1 - \frac{17}{24} = \frac{7}{24}\,.$$

(e) Indicando con Y la variabile aleatoria che conta il numero di piantine con 3 o più foglie tra le 10 del mazzetto, abbiamo $Y \sim B(10, 7/24)$ per quanto visto al punto precedente, e dunque

$$\begin{aligned}\mathbb{P}(Y \geq 2) = 1 - \mathbb{P}(Y < 2) = 1 - \mathbb{P}(Y \leq 1) &= 1 - \left[\sum_{k\in\{0,1\}} \binom{10}{k} \left(\frac{7}{24}\right)^k \left(1 - \frac{7}{24}\right)^{10-k}\right]\\ &= 1 - \left[1\cdot\left(\frac{7}{24}\right)^0\cdot\left(\frac{17}{24}\right)^{10} + 10\cdot\left(\frac{7}{24}\right)^1\cdot\left(\frac{17}{24}\right)^9\right] = 1 - [0.03180 + 0.13093]\\ &= 83.728\%\,.\end{aligned}$$

(f) Se usiamo 45 piantine di basilico, il numero totale di foglie che riusciamo a ottenerne è la v.a. $Z = X_1 + X_2 + \ldots + X_{45}$, in cui X_i è il numero di foglie sull'i-esima piantina, e le v.a. $X_1, X_2, \ldots, X_{45}$ sono i.i.d. con la stessa densità p di X. Per il TLC, la probabilità richiesta è

$$\begin{aligned}\mathbb{P}(Z \geq 80) &= \mathbb{P}\left(\underbrace{\frac{Z - \mathbb{E}[Z]}{\sqrt{\mathrm{Var}(Z)}}}_{\approx N(0,1)} \geq \frac{80 - \mathbb{E}[Z]}{\sqrt{\mathrm{Var}(Z)}}\right) = 1 - \Phi\left(\frac{80 - \mathbb{E}[Z]}{\sqrt{\mathrm{Var}(Z)}}\right) = 1 - \Phi\left(\frac{80 - n\mathbb{E}[X_i]}{\sqrt{n\mathrm{Var}(X_i)}}\right)\\ &= 1 - \Phi\left(\frac{80 - 45\cdot\frac{11}{6}}{\sqrt{45\cdot\frac{53}{36}}}\right) = 1 - \Phi(-0.307) = \Phi(0.307) = 0.62172\\ &= 62.172\%\,.\end{aligned}$$

Dal momento che Z è una v.a. discreta, per ottenere un risultato più preciso si può usare la correzione di continuità, cioè

$$\begin{aligned}\mathbb{P}(Z \geq 80) = \mathbb{P}(Z \geq 79.5) = \ldots &= 1 - \Phi\left(\frac{79.5 - 45\cdot\frac{11}{6}}{\sqrt{45\cdot\frac{53}{36}}}\right) = 1 - \Phi(-0.369) = \Phi(0.369) = 0.64431\\ &= 64.431\%\,.\end{aligned}$$

Problema 2. Un fisico ha appena acquistato una termocoppia nuova per effettuare delle misure di temperatura in laboratorio. Secondo il manuale delle istruzioni, la termocoppia non dovrebbe avere alcun errore sistematico, mentre la sua precisione – intesa come la deviazione standard di una qualsiasi misura effettuata con la termocoppia – dovrebbe essere pari a $1.5°$ C. Il fisico, tuttavia, non si fida completamente di quanto scritto sul manuale, e così, prima di accettare le specifiche dichiarate, decide di metterle alla prova facendo una serie di misure.

La prova consiste nell'usare la termocoppia per rilevare per 30 volte di seguito la temperatura d'ebollizione dell'acqua distillata a pressione standard. Il fisico sa che la temperatura che troverà in una qualsiasi delle 30 misure si può modellizzare con una variabile aleatoria T, per la quale

- $\mathbb{E}[T] = 100° + \delta$, dove δ è l'*errore sistematico* della termocoppia;
- $\mathrm{Var}(T) = \sigma^2$, dove σ è la sua *precisione*.

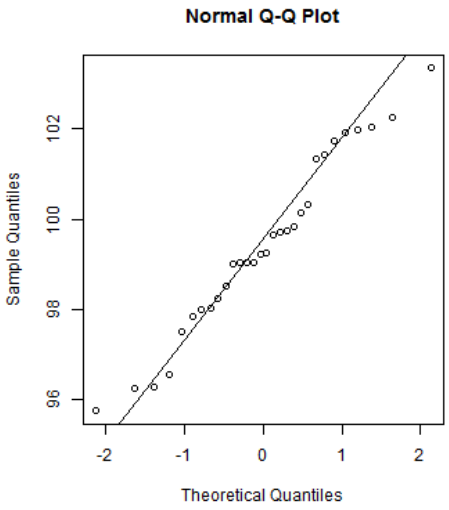
Il fisico raccoglie i risultati delle 30 misure nel vettore `data`, ed elabora questo vettore usando le seguenti righe di codice R, con a fianco il loro output:

```
> data
 [1]  98.991  99.048  96.228 102.054 101.332  99.827  99.030
 [8] 101.443  96.284  97.499 101.934  98.230  96.549 100.327
[15] 101.729  95.745 101.984  99.712  97.851 102.266  99.644
[22] 100.138  98.013 103.349  97.985  99.751  99.045  98.507
[29]  99.208  99.254
> mean(data)
[1] 99.4319
> sd(data)
[1] 1.974782
> shapiro.test(data)

        Shapiro-Wilk normality test

data:  data
W = 0.97088, p-value = 0.5635

> qqnorm(data)
> qqline(data)
>
```



(a) Con un opportuno test al livello di significatività del 5%, stabilite se si può accettare che $\sigma = 1.5°$ C come dichiarato sul manuale della termocoppia, oppure se i dati dimostrano chiaramente il contrario.

(b) Quali sono le ipotesi sul campione del test del punto (a)? Sono rispettate dai dati?

(c) Con un altro test al livello del 5%, decidete se c'è evidenza dai dati che l'errore sistematico δ della termocoppia sia diverso da zero, al contrario di quanto dichiarato invece sul manuale.

(d) Calcolate il p-value del test del punto precedente, o fornite almeno un intervallo in cui esso cade.

(e) Quali sono le ipotesi sul campione del test del punto (c)? Sono compatibili coi dati trovati?

Risultati.

(a) Si richiede di fare un test per le ipotesi statistiche

$$H_0 : \sigma^2 = 1.5^2 =: \sigma_0^2 \qquad \text{contro} \qquad H_1 : \sigma^2 \neq \sigma_0^2 \,.$$

Supponiamo che i 30 dati siano stati estratti da una densità gaussiana. Allora possiamo usare la regola di un test bilatero di livello α per la varianza di un campione gaussiano a media e varianze incognite:

$$\text{“rifiuto } H_0 \text{ se } X_0^2 < \chi^2_{\frac{\alpha}{2}}(n-1) \text{ oppure } X_0^2 > \chi^2_{1-\frac{\alpha}{2}}(n-1)\text{”}\,,$$

dove

$$X_0^2 := \frac{(n-1)S^2}{\sigma_0^2}\,.$$

Dall'output di R ricaviamo $s = 1.974782$, e quindi

$$x_0^2 = \frac{(30-1)\cdot 1.974782^2}{1.5^2} = 50.26362\,,$$

mentre al livello $\alpha = 5\%$

$$\chi^2_{\frac{\alpha}{2}}(n-1) = \chi^2_{0.025}(29) = 16.0471\,, \qquad \chi^2_{1-\frac{\alpha}{2}}(n-1) = \chi^2_{0.975}(29) = 45.7223\,.$$

Poiché

$$50.26362 \notin (16.0471\,,\, 45.7223)\,,$$

la regola del test ci impone di rifiutare H_0, e pertanto di concludere al livello del 5% che $\sigma \neq 1.5°$ (conclusione forte).

(b) L'ipotesi del test precedente è che le 30 misure costituiscano un campione gaussiano. Tale ipotesi è soddisfatta in virtù dell'elevato valore del p-value del test di Shapiro-Wilk (addirittura maggiore del 50%) e della buona aderenza dei quantili empirici alla Q-Q line nel normal Q-Q plot.

(c) Ora si richiede di fare un test per le ipotesi

$$H_0 : \delta = 0 \qquad \text{contro} \qquad H_1 : \delta \neq 0$$

o, equivalentemente,

$$H_0 : \mu = 100 =: \mu_0 \qquad \text{contro} \qquad H_1 : \mu \neq \mu_0\,, \tag{$*$}$$

dove

$$\mu = \mathbb{E}\,[T] = 100 + \delta\,.$$

Poiché abbiamo le realizzazioni delle 30 v.a. $T_1, \ldots, T_{30}$, testiamo direttamente le ipotesi $(*)$ sul valore atteso delle T_i. Come abbiamo visto nel punto precedente, le T_i possono considerarsi gaussiane per il risultato del test di Shapiro-Wilk e per il loro normal Q-Q plot. Tuttavia, la loro varianza σ^2 è incognita, poiché abbiamo visto nel punto (a) che dobbiamo rigettare quanto dichiarato su σ nel manuale. Pertanto, facciamo un T-test per un campione gaussiano a varianza incognita. Con le ipotesi statistiche $(*)$, abbiamo dunque la regola

$$\text{“ rifiuto } H_0 \text{ se } |T_0| > t_{1-\frac{\alpha}{2}}(n-1)\text{”}\,, \tag{$\circ$}$$

dove

$$T_0 = \frac{\bar{X} - \mu_0}{S}\sqrt{n}\,.$$

Coi dati a disposizione, abbiamo la realizzazione

$$t_0 = \frac{99.4319 - 100}{1.974782}\sqrt{30} = -1.57567\,,$$

il cui modulo va confrontato al livello $\alpha = 5\%$ col quantile

$$t_{1-\frac{\alpha}{2}}(n-1) = t_{0.975}(29) = 2.0452\,.$$

Poiché

$$|t_0| = 1.57567 \not> 2.0452\,,$$

dobbiamo accettare H_0 al livello del 5% e concludere che l'errore sistematico è zero, come dichiarato sul manuale.

(d) Il p-value si ricava imponendo l'uguaglianza nella regola di rifiuto $(\circ)$, cioè

$$|t_0| \equiv t_{1-\frac{\alpha}{2}}(n-1)\,.$$

Dalle tavole, vediamo che

$$\begin{aligned} t_{0.90}(29) = 1.3114 < |t_0| = 1.57567 < t_{0.95}(29) = 1.6991 \quad &\Leftrightarrow \quad 0.90 < 1 - \frac{\alpha}{2} < 0.95 \\ &\Leftrightarrow \quad 0.10 < \alpha < 0.20\,. \end{aligned}$$

Pertanto,

$$10\% < p\text{-value} < 20\%\,,$$

che è un intervallo di valori così alti da non permetterci di rifiutare H_0 a nessun livello di significatività ragionevole.

(e) Anche il test del punto (c) si può fare solo se $T_i \sim N(\mu, \sigma^2)$ per ogni i. Tale ipotesi è rispettata sempre in virtù del p-value del test di Shapiro-Wilk e del normal Q-Q plot.

Problema 3. Il signor Bacco vuole studiare come le percentuali di `zucchero` e di `sali` minerali presenti nel mosto d'uva influenzino il grado alcolico (`alcol`) dopo la fermentazione. A tale scopo raccoglie i dati relativi a un certo numero di bottiglie di varie tipologie di vino. Decide di studiare tre diversi modelli di regressione lineare gaussiana, di cui sono riportati gli output di R in Figura 1 e i grafici dei residui in Figura 2.

(a) Scrivere le relazioni ipotizzate nei tre modelli impostati dal signor Bacco.

(b) Quali sono le ipotesi alla base del modello di regressione lineare gaussiana? Per quali dei tre modelli sono verificate? *Giustificare adeguatamente la risposta.*

(c) Scegliere il migliore fra i tre modelli, commentando la significatività dei regressori e indicando la percentuale di variabilità spiegata.

(d) Il signor Bacco vuole avere una stima puntuale della quantità di `alcol` presente in una bottiglia il cui mosto contiene il 20% di `zucchero` e lo 0.3% di `sali`. Quale stima può dare in base ai suoi dati?

(e) Fornire una stima intervallare al 95% per il valore atteso del grado alcolico nella bottiglia del punto precedente, sapendo che $\overline{\texttt{zucchero}} = 22.86\%$ e $\overline{\texttt{sali}} = 0.29\%$.

Risultati.

(a)
- modello 1: $\texttt{alcol} = \beta_0 + \beta_1 \texttt{zucchero} + \epsilon, \qquad \epsilon \sim \mathcal{N}(0, \sigma^2),$
- modello 2: $\texttt{alcol} = \beta_0 + \beta_1 \texttt{sali} + \epsilon, \qquad \epsilon \sim \mathcal{N}(0, \sigma^2),$
- modello 3: $\texttt{alcol} = \beta_0 + \beta_1 \texttt{zucchero} + \beta_2 \texttt{sali} + \epsilon, \qquad \epsilon \sim \mathcal{N}(0, \sigma^2).$

(b) Le ipotesi alla base del modello sono che i residui siano normali e omoschedastici. L'ipotesi di omoschedasticità è soddisfatta per tutti i modelli, come si può vedere dagli scatterplot dei residui che non presentano particolari pattern. L'ipotesi di normalità è verificata per i modelli 1 e 3 dato che i quantili teorici e empirici nel qq-plot seguono l'andamento lineare e il p-value dello shapiro-test è maggiore di 0.05. Nel modello 2 il qq-plot non presenta un andamento lineare e inoltre il p-value dello shapiro-test è inferiore a 0.05, quindi i residui non possono essere considerati normali.

(c) Il modello migliore tra i tre risulta essere il modello 1. Infatti, il modello 2 non soddisfa le ipotesi sui residui e di conseguenza va scartato. Il modello 3 presenta un regressore non significativo (`sali`) dato che il test sul relativo coefficiente ha un p-value di 0.10169; inoltre ha un R^2 praticamente identico a quello del modello 1, di conseguenza si preferisce scegliere il modello più semplice.

(d) Scegliendo il modello 1 la previsione è:

$$\widehat{\texttt{alcol}} = -1.9388 + 0.6223 * 20 = 10.5072.$$

(e) Dobbiamo fornire un intervallo di confidenza per la media di una nuova osservazione. Si ha:

$$s_{xx} = \frac{\hat{\sigma}^2}{\mathrm{se}^2(\beta_1)} = \frac{0.7067^2}{0.0255^2} = 768.05.$$

Usando la formula per l'intervallo di confidenza abbiamo:

$$\begin{aligned} IC_{0.95} &= \left(\hat{\beta}_0 + \hat{\beta}_1 x^* \pm t_{1-\frac{\alpha}{2}}(n-2) \sqrt{\hat{\sigma}^2 \left(\frac{1}{n} + \frac{(x^* - \bar{x})^2}{s_{xx}} \right)} \right) = \\ &= \left(10.5072 \pm 2.0211 \sqrt{0.7067^2 \left(\frac{1}{42} + \frac{(20 - 22.86)^2}{768.05} \right)} \right) = \\ &= (10.24206,\ 10.77234) \end{aligned}$$

```
> summary(modello1)

Call:
lm(formula = alcol ~ zucchero)

Residuals:
     Min       1Q   Median       3Q      Max
-1.66946 -0.46924  0.05446  0.42389  1.27320

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -1.9388     0.5929   -3.27  0.00222 **
zucchero      0.6223     0.0255   24.41  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7067 on 40 degrees of freedom
Multiple R-squared:  0.9371,    Adjusted R-squared:  0.9355
F-statistic: 595.7 on 1 and 40 DF,  p-value: < 2.2e-16

> summary(modello2)

Call:
lm(formula = alcol ~ sali)

Residuals:
    Min      1Q  Median      3Q     Max
-3.8857 -2.6056 -0.0805  2.2925  4.5008

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   10.909      0.734  14.862   <2e-16 ***
sali           4.691      2.074   2.262   0.0292 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.653 on 40 degrees of freedom
Multiple R-squared:  0.1134,    Adjusted R-squared:  0.09122
F-statistic: 5.115 on 1 and 40 DF,  p-value: 0.02922

> summary(modello3)

Call:
lm(formula = alcol ~ zucchero + sali)

Residuals:
     Min       1Q   Median       3Q      Max
-1.72693 -0.52763  0.09633  0.37340  1.43314

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -1.9339     0.5800  -3.334  0.00188 **
zucchero      0.6099     0.0260  23.455  < 2e-16 ***
sali          0.9447     0.5636   1.676  0.10169
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6912 on 39 degrees of freedom
Multiple R-squared:  0.9413,    Adjusted R-squared:  0.9383
F-statistic: 312.7 on 2 and 39 DF,  p-value: < 2.2e-16
```

Figura 1: Summary dei tre modelli stimati.

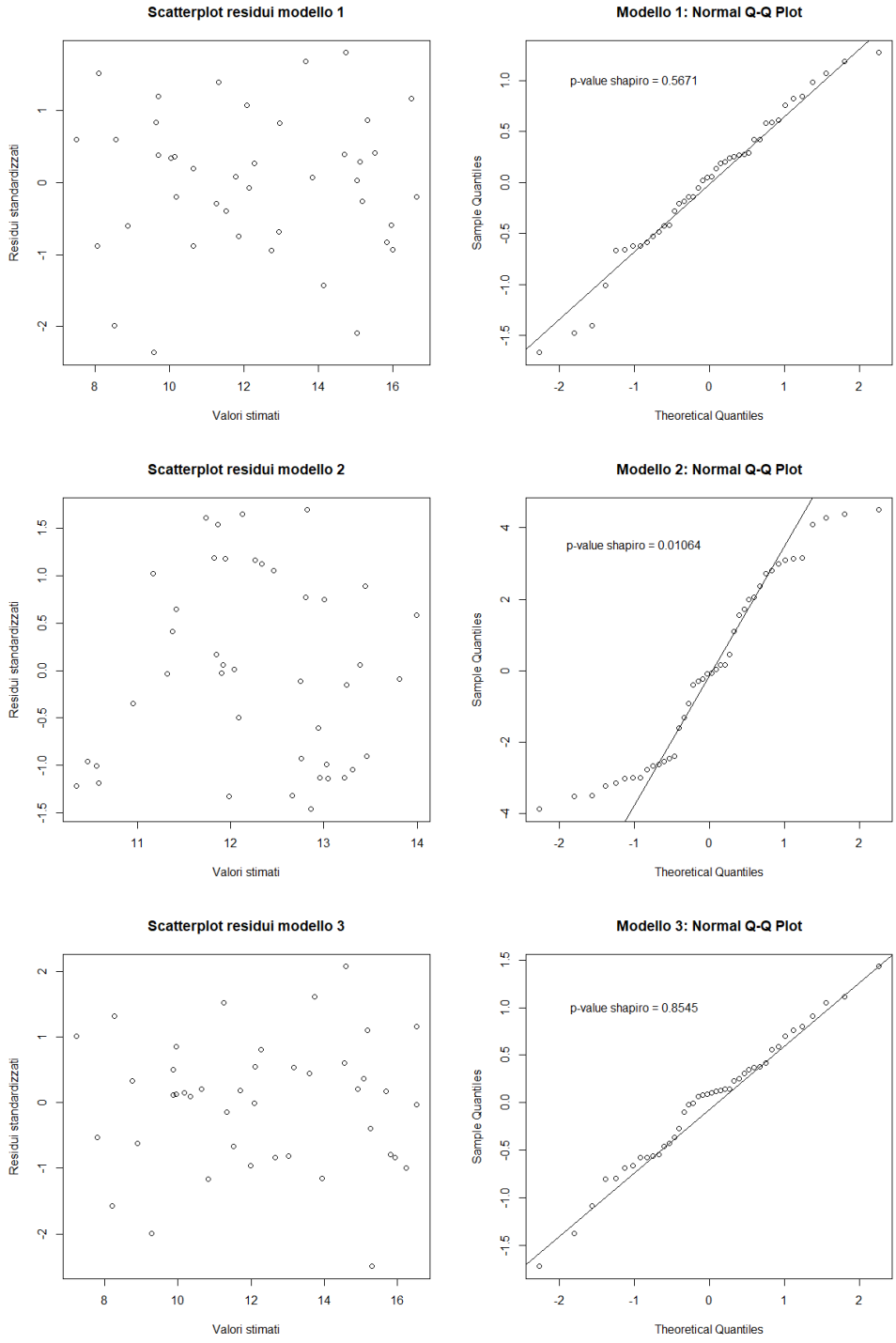


Figura 2: Scatterplot e qq-plot dei residui dei tre modelli stimati.