Esercizi Unità 4

Analisi dei dati 2022/23

Cristiano Varin

1. L'azienda di elettronica di consumo Simsang vuole valutare due diverse strategie per promuovere le vendite di un nuovo televisore. La prima strategia consiste nel vendere la televisione con in omaggio un abbonamento di tre anni ad un servizio di streaming di serie TV e film. La seconda strategia, invece, consiste nel vendere il televisore con uno sconto. Per valutare quale delle due strategie sia più gradita, viene intervistato un campione casuale di 200 clienti. Il numero di clienti che dichiarano di preferire il televisore con lo sconto è stato 110.

   (a) Si calcoli un intervallo di confidenza al 90% per la probabilità di preferire l'offerta della televisione con l'abbonamento al servizio di streaming piuttosto che la televisione con lo sconto.

   (b) Si verifichi se vi sia una differenza nel livello di gradimento verso le due strategie di promozione al livello di significatività del 5%.

   (c) Dopo mezzo anno viene intervistato un secondo campione casuale di 250 clienti. Il numero di clienti in questo secondo campione che dichiara di preferire il televisore con lo sconto è 105. Questo risultato conferma che i clienti ora preferiscono maggiormente l'abbonamento di quanto non lo preferissero prima?

   Soluzione

   (a) Indichiamo con $X_1, \ldots, X_{200}$ il campione casuale di clienti intervistati. Le osservazioni X_i sono variabili binarie che assumono il valore 1 se viene preferito il televisore con l'abbonamento al servizio di streaming e 0 se, invece, viene preferito il televisore con lo sconto. Indichiamo con p la probabiltà di preferire l'abbonamento allo sconto. Un intervallo di confidenza per p di livello approssimativamente pari a 90% è

   $$\hat{p} \pm z_{0.05}\sqrt{\frac{\hat{p}(1-\hat{p})}{200}},$$

   dove $\hat{p} = \bar{X}$ e $z_{0.05} = 1.64$. I dati campionari danno $\hat{p} = 90/200 = 0.45$, per cui l'intervallo vale

   $$0.45 \pm 1.64\sqrt{\frac{0.45(0.55)}{200}} = [0.39, 0.51].$$

   (b) La domanda richiede di verificare il sistema d'ipotesi con alternativa bilaterale

   $$\mathrm{H}_0 : p = 0.5 \quad \text{contro} \quad \mathrm{H}_A : p \neq 0.5$$

   poiché se non vi è differenza fra le due promozioni allora la probabilità di preferire una delle due promozioni è pari a 0.5. Sotto H_0, lo stimatore $\hat{p}$ ha distribuzione approssimativamente normale con media 0.5 e varianza $0.5(1-0.5)/200$. Possiamo, quindi,

valutare le ipotesi con la statistica test Z

$$Z = \frac{\sqrt{200}(\hat{p} - 0.5)}{0.5}.$$

La statistica Z ha distribuzione approssimativamente normale standard sotto H_0. La regione di rifiuto con livello di significatività approssimativamente del 5% è

$$\mathcal{R} = \{(-\infty, -z_{0.025}] \cup [z_{0.025}, +\infty)\} = \{(-\infty, -1.96] \cup [1.96, +\infty)\}.$$

Il valore osservato della statistica Z è $z = -1.41$, quindi *non* possiamo rifiutare l'ipotesi nulla che le due promozioni siano gradite allo stesso modo dai clienti.

Potevamo giungere a questa conclusione senza svolgere calcoli perché nella soluzione del precedente quesito abbiamo visto che un intervallo di confidenza del 90% contiene il valore sotto ipotesi nulla $p_0 = 0.5$ e quindi a maggior ragione questo valore sarà contenuto nelll'intervallo di confidenza di livello 95% ad indicare che non possiamo rifiutare l'ipotesi nulla ad un livello di significatività del 5%.

(c) Indichiamo con p_1 la probabilità di preferire l'abbonamento allo sconto al tempo della prima rilevazione e con p_2 la stessa probabilità dopo mezzo anno. Vogliamo valutare il sistema d'ipotesi con alternativa unilaterale sinistra

$$H_0 : p_1 = p_2 \quad \text{contro} \quad H_A : p_1 < p_2,$$

che possiamo anche scrivere come

$$H_0 : \theta = 0 \quad \text{contro} \quad H_A : \theta < 0,$$

dove $\theta = p_1 - p_2$. Consideriamo lo stimatore non distorto e asintoticamente normale $\hat{\theta} = \hat{p}_1 - \hat{p}_2$, dove $\hat{p}_1$ e $\hat{p}_2$ sono le proporzioni campionarie di clienti che preferiscono l'abbonamento nelle due rilevazioni. Valutiamo le ipotesi con la statistica test

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1 - \hat{p})\left(\frac{1}{200} + \frac{1}{250}\right)}},$$

dove $\hat{p}$ è la stima "pooled" ottenuta utilizzando i dati di entrambe le rilevazioni, ovvero la proporzione complessiva di clienti che ha dichiarato di preferire l'abbonamento. La statistica Z ha distribuzione approssimativamente normale standard sotto H_0. Nel testo dell'esercizio non è specificato il livello di significatività, quindi possiamo sceglierlo a nostro piacimento. Per esempio, la regione di rifiuto con livello di significatività approssimativamente del 5% è

$$\mathcal{R} = (-\infty, -z_{0.05}] = (-\infty, -1.64].$$

Con i dati osservati abbiamo che $\hat{p}_1 = 0.45$, $\hat{p}_2 = 145/250 = 0.58$ e $\hat{p} = (90 + 145)/(200 + 250) = 0.52$. Il valore osservato della statistica Z è,

$$z = \frac{0.45 - 0.58}{\sqrt{0.52(1 - 0.52)\left(\frac{1}{200} + \frac{1}{250}\right)}} = -2.74,$$

quindi rifiutiamo l'ipotesi nulla al livello di significatività approssimativamente del 5% e concludiamo che sia aumentato nel tempo il grado di preferenza verso l'offerta dell'abbonamento al servizio di streaming rispetto all'offerta dello sconto.

2. L'installazione di un software su un computer richiede un ammontare di tempo casuale. Un tecnico informatico installa il software in un laboratorio composto da 24 computer con un tempo medio di installazione per singolo computer di 4.6 minuti con deviazione standard di 2.1 minuti. Assumendo che il tempo di installazione sia normalmente distribuito si risponda ai seguenti quesiti.

   (a) Si calcoli un intervallo di confidenza al 95% per il tempo medio di installazione del software su un singolo computer.

   (b) Dopo sei mesi viene rilasciata una nuova versione del software che viene installata in un secondo laboratorio che consiste di 18 computer dello stesso modello dei computer del primo laboratorio. Il tempo medio per installare il software su un singolo computer è pari a 4.1 minuti con una deviazione standard di 1.9 minuti. Il tempo di installazione della nuova versione software è davvero più veloce di quello della precedente versione del software? Si risponda alla domanda ad un livello di significatività del 1%.

Soluzione

(a) L'intervallo di confidenza al 95% per il tempo di installazione medio su un singolo computer è

$$4.6 \pm t_{0.025} \frac{2.1}{\sqrt{24}},$$

dove $t_{0.025}$ è il quantile di posizione 0.975 della distribuzione T di Student con 23 gradi di libertà. Con **R** troviamo che $t_{0.025} = 2.07$, per cui l'intervallo di confidenza è $[3.71, 5.49]$ minuti.

(b) Dobbiamo valutare il sistema d'ipotesi

$$\mathrm{H}_0 : \mu_v = \mu_n \quad \text{contro} \quad \mathrm{H}_A : \mu_v > \mu_n,$$

dove μ_v and μ_n sono i tempo medi di installazione del software su un singolo computer con la vecchia e il nuova versione del software. Possiamo valutare le ipotesi con la statistica T

$$T = \frac{\bar{X}_v - \bar{X}_n}{\sqrt{\left(\frac{S_v^2}{24} + \frac{S_n^2}{18}\right)}}.$$

Sotto l'ipotesi nulla, la statistica T ha una distribuzione approssimativamente T di Student con

$$\nu = \frac{\left(\frac{2.1^2}{24} + \frac{1.9^2}{18}\right)^2}{\frac{2.1^4}{24^2(23)} + \frac{1.9^4}{18^2(17)}} = 38.52$$

gradi di libertà. La regione di rifiuto del test è

$$\mathcal{R} = [t_{0.01}, +\infty),$$

dove $t_{0.01}$ è il quantile di posizione 0.99 della distribuzione T di Student con 38.52 gradi di libertà. Usando **R** troviamo che $t_{0.01} = 2.43$. Il valore osservato della statistica T è

$$t = \frac{4.6 - 4.1}{\sqrt{\left(\frac{2.1^2}{24} + \frac{1.9^2}{18}\right)}} = 0.81,$$

per cui non possiamo rifiutare l'ipotesi nulla al livello di significatività del 1%.

3. Ogni giorno i programmatori che lavorano in una software house trovano un certo numero di errori casuali nei loro codici e li correggono. In un campione casuale di 32 giorni sono stati trovati complessivamente 374 errori. Siccome questo numero di errori è considerato troppo elevato dal manager dell'azienda, viene deciso di far seguire a tutti i programmatori dell'azienda un corso per migliorare la loro capacità di concentrazione durante la stesura del codice. Successivamente al corso, in un campione casuale di 37 giorni vengono trovati complessivamente 390 errori.

   Supponendo che il numero di errori segua una distribuzione di Poisson, si risponda ai seguenti quesiti:

   (a) Possiamo ritenere che il numero medio di errori giornalieri prima di seguire il corso fosse significativamente superiore a 10? Si risponda al quesito ad un livello di significatività del 5%.

   (b) Sulla base dei campioni osservati, possiamo affermare che il corso ha portato ad una riduzione del numero medio di errori giornalieri? Si risponda al quesito ad un livello di significatività del 1%.

   Soluzione

   (a) Indichiamo con λ il numero medio di errori. Dobbiamo valutare il sistema d'ipotesi con alternativa unilaterale destra

   $$\mathrm{H}_0 : \lambda = 10 \quad \text{contro} \quad \mathrm{H}_A : \lambda > 10.$$

   Procediamo con la statistica test Z

   $$Z = \frac{\sqrt{32}(\bar{X} - 10)}{\sqrt{10}}$$

   che ha approssimativamente una distribuzione normale standard sotto l'ipotesi nulla. La regione di rifiuto con un livello di significatività del 5% è

   $$\mathcal{R} = [1.645, +\infty).$$

   Con i dati osservati abbiamo $\bar{x} = 374/32 = 11.69$ e il valore osservato della statistica test è $z = 3.01$ per cui rifiutiamo l'ipotesi nulla al livello di significatività del 5%.

   (b) Indichiamo con λ_p il numero medio di errori prima del corso e con λ_d il numero medio di errori dopo il corso. Dobbiamo valutare il sistema d'ipotesi con alternativa unilaterale destra

   $$\mathrm{H}_0 : \lambda_p = \lambda_d \quad \text{contro} \quad \mathrm{H}_A : \lambda_p > \lambda_d.$$

   Procediamo con la statistica test Z

   $$Z = \frac{(\bar{X} - \bar{Y})}{\sqrt{\frac{\bar{X}}{32} + \frac{\bar{Y}}{37}}},$$

   dove $\bar{X}$ è la media campionaria del numero di errori prima del corso e $\bar{Y}$ è la media campionaria del numero di errori dopo il corso. La statistica Z ha approssimativamente

una distribuzione normale standard sotto l'ipotesi nulla. La regione di rifiuto con livello di significatività del 1% è

$$\mathcal{R} = [2.32, +\infty).$$

Con i dati osservati abbiamo $\bar{x} = 11.69$, $\bar{y} = 390/37 = 10.54$, e il valore osservato della statistica test $z = 1.43$ per cui non possiamo rifiutare l'ipotesi nulla ovvero concludere che il corso abbia portato ad una riduzione statisticamente significativa del numero di errori.

4. Sia $X_1, \ldots, X_{200}$ un campione casuale semplice da una distribuzione discreta con funzione di probabilità

$$\Pr(X = x; \theta) = (1 - e^{\theta})^x e^{\theta}, \quad x = 0, 1, \ldots, \ \theta < 0.$$

Si consideri il sistema d'ipotesi

$$\mathrm{H}_0 : \theta = -2 \quad \text{vs} \quad \mathrm{H}_A : \theta \neq -2\,.$$

(a) Si proponga una statistica Z per valutare il sistema d'ipotesi.

(b) Si proponga una regola per valutare le ipotesi che garantisca un livello di significatività approssimativamente pari al 1%.

(c) Si valuti il sistema d'ipotesi nel caso di un campione che ha dato $\sum_{i=1}^{200} x_i = 1025$.

(d) Si valuti ora il sistema d'ipotesi unilaterale sinistro

$$\mathrm{H}_0 : \theta = -2 \quad \text{vs} \quad \mathrm{H}_A : \theta < -2$$

ad un livello di significatività approssimativamente pari al 1%.

Soluzione

(a) La log-verosimiglianza à

$$\ell(\theta) = \log(1 - e^{\theta}) \sum_{i=1}^{n} X_i + n\theta.$$

La funzione punteggio è

$$\ell'(\theta) = -\frac{e^{\theta}}{1 - e^{\theta}} \sum_{i=1}^{n} X_i + n.$$

Il punto stazionario della log-verosimiglianza è

$$\hat{\theta} = -\log(1 + \bar{X})$$

e corrisponde allo stimatore di massima verosimiglianza poiché

$$\ell''(\theta) = -\frac{e^{\theta}}{(1 - e^{\theta})^2} \sum_{i=1}^{n} X_i < 0, \text{ per qualsiasi } \theta.$$

L'informazione osservata è

$$J(\theta) = -\ell''(\theta) = \frac{e^{\theta}}{(1 - e^{\theta})^2} \sum_{i=1}^{n} X_i.$$

L'errore standard stimato di $\hat{\theta}$ è approssimativamente pari a

$$\begin{aligned}\text{se}(\hat{\theta}) &\approx \sqrt{J(\hat{\theta})^{-1}} \\ &= \sqrt{\frac{(1-e^{\hat{\theta}})^2}{e^{\hat{\theta}}\, n\, \bar{X}}} \\ &= \sqrt{\frac{1-e^{\hat{\theta}}}{n}}.\end{aligned}$$

Nell'ultimo passaggio abbiamo usato il fatto che $\bar{X} = (1-e^{\hat{\theta}})/e^{\hat{\theta}}$.

Possiamo valutare le ipotesi con la statistica Z

$$Z = \frac{\sqrt{n}(\hat{\theta}+2)}{\sqrt{1-e^{\hat{\theta}}}}.$$

Sotto l'ipotesi nulla, la statistica Z ha distribuzione approssimativamente normale standard.

(b) La regione di rifiuto del test che assicura un livello di significatività approssimativamente pari al 1% è

$$\mathcal{R} = \{(-\infty, -z_{0.005}] \cup [z_{0.005}, +\infty)\} = \{(-\infty, -2.58] \cup [2.58, +\infty)\}.$$

(c) La stima di massima verosimiglianza con i dati osservati è $\hat{\theta} = -1.81$ e il valore osservato della statistica Z è $z = 2.94$. Quindi, rifiutiamo l'ipotesi nulla al livello di significatività del 1%.

(d) Non possiamo rifiutare l'ipotesi nulla. Infatti, nel caso dell'ipotesi alternativa unilaterale la regione di rifiuto diventa

$$\mathcal{R} = (-\infty, -z_{0.01}] = (-\infty, -2.33].$$

5. Sia $X_1, \ldots, X_{50}$ un campione casuale semplice dalla distribuzione continua

$$f(x;\theta) = \frac{\theta}{2\sqrt{x}} \exp\left(-\theta\sqrt{x}\right), \; x > 0, \; \theta > 0.$$

Supponendo che il campione osservato abbia dato $\sum_{i=1}^{50} \sqrt{x_i} = 37.9$:

(a) Si calcoli la stima di massima verosimiglianza.

(b) Si calcoli un intervallo di confidenza per θ di livello 99%.

(c) Si verifichi il sistema d'ipotesi

$$\text{H}_0 : \theta = 1.9 \quad \text{contro} \quad \text{H}_A : \theta \neq 1.9.$$

Soluzione

(a) La log-verosimiglianza è

$$\ell(\theta) = n \log \theta - \theta \sum_{i=1}^{n} \sqrt{X_i},$$

con corrispondente funzione punteggio

$$\ell'(\theta) = \frac{n}{\theta} - \sum_{i=1}^{n} \sqrt{X_i}.$$

La soluzione dell'equazione di verosimiglianza

$$\hat{\theta} = n / \sum_{i=1}^{n} \sqrt{X_i}$$

è lo stimatore di massima verosimiglianza poiché

$$\ell''(\theta) = -\frac{n}{\theta^2} < 0, \text{ per qualsiasi } \theta.$$

Con i dati osservati otteniamo la stima di massima verosimiglianza $\hat{\theta} = 1.32$.

(b) Un intervallo di confidenza per θ di livello approssimato del 99% è

$$\hat{\theta} \pm z_{0.005} \frac{\hat{\theta}}{\sqrt{n}}.$$

Con **R** troviamo che $z_{0.005} = 2.58$ per cui l'intervallo di confidenza corrispondente al campione osservato è $[0.84, 1.80]$.

(c) La domanda non specifica il livello di significatività a cui condurre la verifica d'ipotesi. In questo caso è conveniente scegliere il livello di significatività del 1%. Infatti, siccome l'intervallo di confidenza al 99% non contiene il valore 1.9, possiamo concludere che rifiutiamo l'ipotesi nulla al livello di significatività del 1%.

6. I dati contenuti nel file **monet.csv** riguardano il prezzo di vendita in milioni di dollari di diversi dipinti di Claude Monet. Oltre al prezzo di vendita (**price**) e l'identificativo numerico del dipinto (**id**), sono disponibili informazioni circa le dimensioni del dipinto (**height** e **width** in pollici), se il dipinto è firmato da Monet o meno (**signed**) e la casa d'aste in cui il dipinto è stato venduto (**house** con le tre case identificate da diversi numeri).

(a) Si valuti la normalità dei prezzi di vendita dei dipinti.

(b) Si valuti la normalità dei prezzi di vendita dei dipinti distinti a seconda se siano stati firmati o no da Claude Monet.

(c) Si rivalutino i due punti precedenti dopo aver trasformato i dati su scala logaritmica.

Soluzione

(a) Leggiamo i dati in **R**:

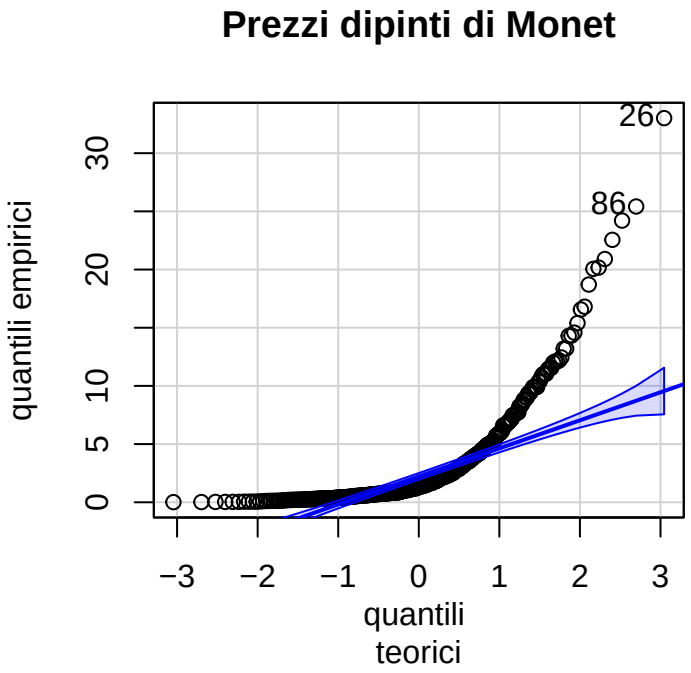
```
R> monet <- read.csv("monet.csv")
```

Carichiamo la libreria **car**:

```
R> library("car")
```

Vediamo il grafico quantile-quantile dei prezzi dei dipinti:

```
R> qqPlot(monet$price, ylab = "quantili empirici", xlab = "quantili
teorici", main = "Prezzi dipinti di Monet")
## [1] 26 86
```



Il grafico mostra in modo estremamente chiaro che i prezzi dei dipinti non si possono ritenere normalmente distribuiti.

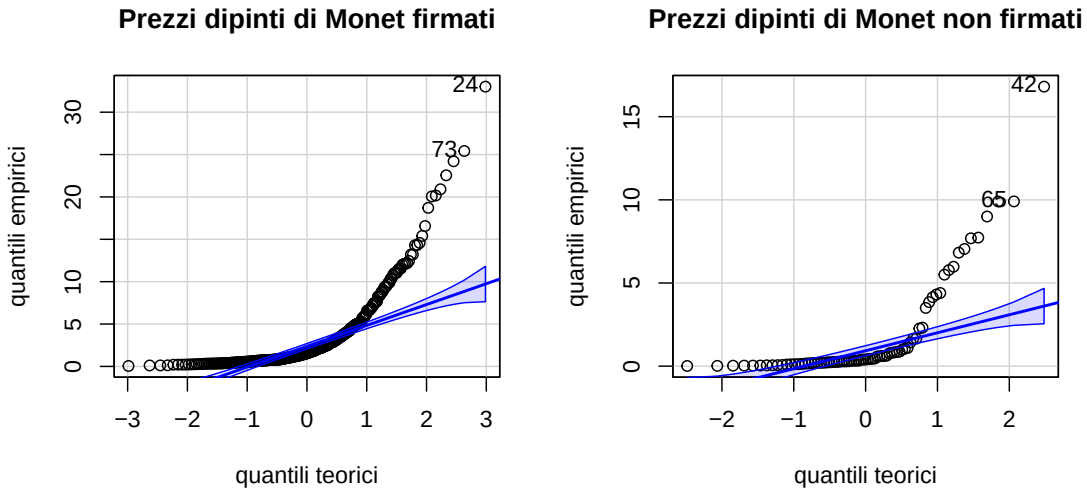
(b) Valutiamo la normalità dei prezzi a seconda se siano stati firmati o no da Claude Monet considerando i due sottogruppi di dipinti:

```
R> x <- with(monet, price[signed == 1])
R> y <- with(monet, price[signed == 0])
```

Vediamo i grafici quantile-quantile per i due gruppi:

```
R> ## il seguente comando divide la finestra grafica in una riga e due
R> ## colonne in modo da poter disegnare i due grafici appaiati
R> par(mfrow = c(1, 2))
R> qqPlot(x, ylab = "quantili empirici", xlab = "quantili teorici",
+ main = "Prezzi dipinti di Monet firmati")
## [1] 24 73
R> qqPlot(y, ylab = "quantili empirici", xlab = "quantili teorici",
+ main = "Prezzi dipinti di Monet non firmati")
## [1] 42 65
R> ## ritorniamo ad una finestra grafica non divisa, ovvero formata da una
```

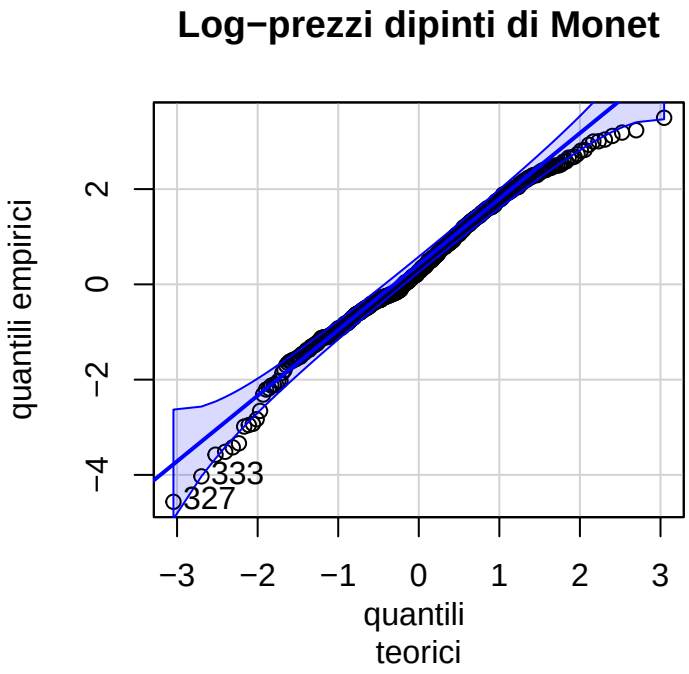
```
R> ## riga e una colonna
R> par(mfrow = c(1, 1))
```



Anche i grafici quantile-quantile dei prezzi distinti a seconda della presenza della firma mostrano un'andamento fortemente non normale.

(c) Ripetiamo i grafici precedenti su scala logartimica. Grafico di tutti i prezzi di vendita:

```
R> qqPlot(log(monet$price), ylab = "quantili empirici", xlab = "quantili
teorici", main = "Log-prezzi dipinti di Monet")
## [1] 327 333
```



La trasformazione logaritmica ha decisamente migliorato la situazione sebbene la disposizione dei punti sulle due code non sia soddisfacente.
Infine, vediamo i grafici quantile-quantile dei prezzi a seconda se i dipinti siano firmati o meno:

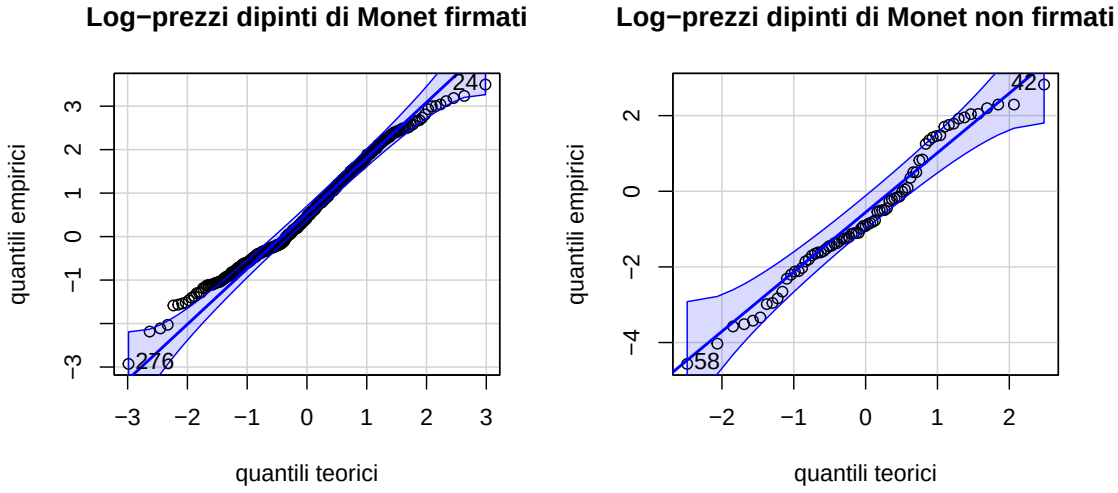
```
R> par(mfrow = c(1, 2))
R> qqPlot(log(x), ylab = "quantili empirici", xlab = "quantili teorici",
+ main = "Log-prezzi dipinti di Monet firmati")

## [1] 276  24

R> qqPlot(log(y), ylab = "quantili empirici", xlab = "quantili teorici",
+ main = "Log-prezzi dipinti di Monet non firmati")

## [1] 58 42

R> par(mfrow = c(1, 1))
```



Il grafico relativo ai dipinti non firmati ci permette di concludere che non possiamo escludere che i prezzi su scala logaritmica siano normalmente distribuiti, mentre per i dipinti firmati gli andamenti sulla due code, e in particolare su quella di sinistra, non sono del tutto soddisfacenti.

7. Utilizzando i dati contenuti nel file **monet.csv**:

   (a) Si valuti se il prezzo di vendita dei dipinti firmati da Claude Monet sia significativamente più alto del prezzo di vendita dei dipinti attribuiti a Claude Monet seppur non firmati dal pittore.

   (b) Si valuti se la proporzione di dipinti firmati che sono stati battuti dalla prima casa d'aste (**house == 1**) sia diversa da quella dei dipinti firmati che sono stati battuti dalla seconda casa d'aste (**house == 2**).

Soluzione

(a) Sulla base dei risultati dell'esercizio precedente procediamo con l'analisi dei dati sulla scala logaritmica e valutiamo il sistema d'ipotesi:

$$\mathrm{H}_0 : \mu_x = \mu_y \quad \text{contro} \quad \mathrm{H}_A : \mu_x > \mu_y,$$

dove μ_x indica la media del logaritmo del prezzo dei dipinti firmati da Monet e μ_y indica la media del logaritmo del prezzo dei dipinti non firmati ma comunque attribuiti a Monet.

Leggiamo i dati in **R**:

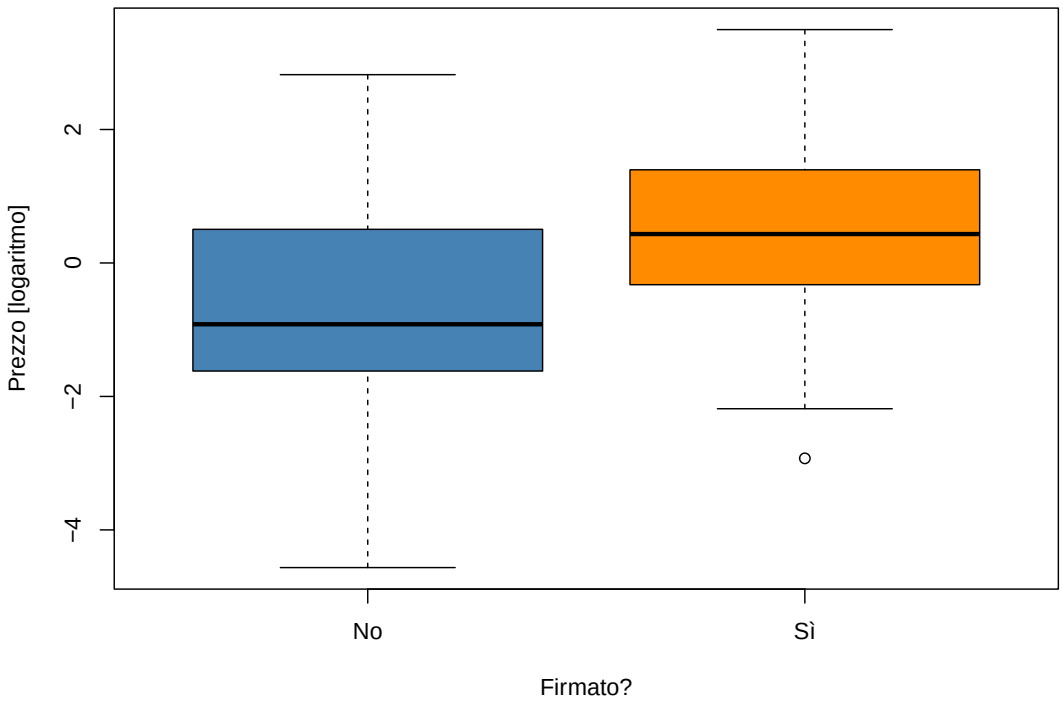
```
R> monet <- read.csv("monet.csv")
```

Estraiamo la variabile binaria che indica se un dipinto sia firmato o meno:

```
R> firmato <- as.factor(monet$signed)
R> levels(firmato) <- c("No", "Sì")
R> head(firmato)
## [1] Sì Sì No Sì Sì Sì
## Levels: No Sì
```

Visualizziamo le distribuzioni delle due variabili:

```
R> boxplot(log(monet$price) ~ firmato, xlab = "Firmato?",
+ ylab = "Prezzo [logaritmo]", col = c("steelblue", "darkorange"))
```



Il grafico mostra che i prezzi di vendita (su scala logaritmica) sono più alti per i dipinti firmati. Valutiamo ora il sistema d'ipotesi per vedere se la differenza osservata è statisticamente significativa. Per prima cosa costruiamo i vettori che contengono i prezzi su scala logaritmica distinti a seconda della firma:

```
R> x <- log(monet$price[firmato == "Sì"])
R> y <- log(monet$price[firmato == "No"])
```

Procediamo con un test di tipo Z:

```
R> library("BSDA")
R> z.test(x, y, sigma.x = sd(x) , sigma.y = sd(y), alternative = "greater")

##
##  Two-sample z-Test
##
## data:  x and y
## z = 5.9926, p-value = 1.032e-09
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  0.8796511        NA
## sample estimates:
##  mean of x  mean of y
##  0.5498502 -0.6625911
```

Il valore osservato della statistica Z è $z = 5.99$ che corrisponde ad un livello di significatività osservato pressoché nullo ad indicare un'evidenza molto forte contro l'ipotesi nulla.

(b) Il secondo quesito richiede di valutare il sistema d'ipotesi

$$\mathrm{H}_0 : p_X = p_Y \quad \text{contro} \quad \mathrm{H}_A : p_X \neq p_Y,$$

dove p_X indica la proporzione di dipinti firmati che sono stati battuti dalla prima casa d'aste e p_Y la proporzione di dipinti firmati che sono stati battuti dalla seconda casa d'aste. Costruiamo le variabili binarie corrispondenti alle due case d'asta:

```
R> x <- with(monet, signed[house == 1])
R> y <- with(monet, signed[house == 2])
```

Le proporzioni camponarie dei dipinti firmati battuti dalle due case d'aste sono:

```
R> mean(x)

## [1] 0.8324607

R> mean(y)

## [1] 0.8604651
```

Valutiamo il sistema d'ipotesi con test Z:

```
R> sd.x <- sqrt(mean(x) * (1 - mean(x)))
R> sd.y <- sqrt(mean(y) * (1 - mean(y)))
R> z.test(x, y, sigma.x = sd.x, sigma.y = sd.y)

##
##  Two-sample z-Test
##
## data:  x and y
## z = -0.78011, p-value = 0.4353
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.09836273  0.04235396
## sample estimates:
## mean of x mean of y
```

```
## 0.8324607 0.8604651
```

Il valore osservato della statistica Z è $z = -0.78$ che corrisponde ad un livello di significatività osservato pari a 0.44. Non possiamo, quindi, rifiutare l'ipotesi nulla.

8. Utilizzando ancora una volta i dati relativi ai 23 lanci dello Space Shuttle Challenger precedenti alla sua esplosione contenuti nel file **challenger.csv**, si verifichi l'ipotesi che temperature basse siano associate alla rottura degli *o-ring*.

Soluzione. Possiamo rispondere al quesito valutando se la temperatura media dei lanci in cui sono avvenute rotture dei *o-ring* sia significativamente più bassa della temperatura media dei lanci in cui *non* sono avvenute rotture dei *o-ring*, ovvero valutare il sistema d'ipotesi

$$\mathrm{H}_0 : \mu_X = \mu_Y \quad \text{contro} \quad \mathrm{H}_A : \mu_X < \mu_Y,$$

dove μ_X è la temperatura media dei lanci con rottura di *o-ring* mente μ_Y è la temperatura media dei lanci senza rottura di *o-ring*.

Leggiamo i dati:

```
R> challenger <- read.csv("challenger.csv")
R> head(challenger)

##   failure temperature
## 1   FALSE          66
## 2    TRUE          70
## 3   FALSE          69
## 4   FALSE          68
## 5   FALSE          67
## 6   FALSE          72
```

Creiamo le variabili che contengono le temperature dei due gruppi di lanci:

```
R> x <- with(challenger, temperature[failure])
R> y <-  with(challenger, temperature[!failure])
R> ## "!" e' l'operatore logico "non"
```

Verifichiamo la normalità dei due gruppi di dati:

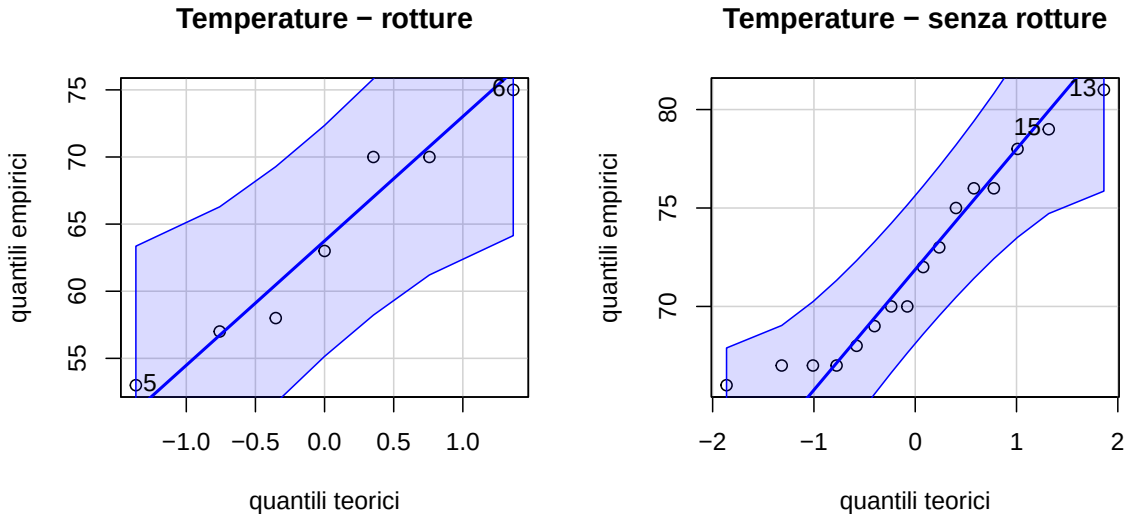
```
R> par(mfrow = c(1, 2))
R> qqPlot(x, ylab = "quantili empirici", xlab = "quantili teorici",
+ main = "Temperature - rotture")

## [1] 6 5

R> qqPlot(y, ylab = "quantili empirici", xlab = "quantili teorici",
+ main = "Temperature - senza rotture")

## [1] 13 15

R> par(mfrow = c(1, 1))
```



I grafici quantile-quantile suggeriscono che l'assunzione di normalità non può essere rifiutata. Viste le numerosità campionarie valutiamo le ipotesi con un test T a due campioni:

```
R> t.test(x, y, alternative = "less")

##
##  Welch Two Sample t-test
##
## data:  x and y
## t = -2.5387, df = 7.9166, p-value = 0.01753
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##       -Inf -2.241649
## sample estimates:
## mean of x mean of y
##  63.71429  72.12500
```

Il valore osservato della statistica T è $t = -2.54$ con un corrispondente livello di significatività osservato pari a 0.018 che indica una debole evidenza contro l'ipotesi nulla.

9. Si svolgano i conti dell'esempio sui test per dati appaiati discusso nelle ultime due pagine dell'Unità 4. L'esempio riguarda la valutaziane se sia davvero conveniente acquistare i libri su **Amazon.com** piuttosto che in comuni librerie. Di seguito sono riportati i prezzi medi di vendita in dollari americani di un campione casuale di 15 manuali universitari su **Amazon.com** e in comuni librerie[1].

[1]Dataset `Textbook` tratto da Levine, Krehbiel & Berenson (2010). Statistica. Pearson.

Libro	Libreria	Amazon
Principles of Microeconomics	120	101.22
Calculus: Early Transcendentals	137.5	115.33
Exploring Wine	65	37.05
Manual de Gramatica	82.75	71.36
Deviant Behavior	90	83
Modern Architecture Since 1900	39.95	26.37
Rise of Christianity	40	26.4
Commercial Banking	120	108.99
A Romance of a Republic	25	14.99
Chemistry in Context	133.75	102.3
Universal Principles of Design	40	26.4
In Mixed Company	79.5	68.76
International Marketing	154.75	126.15
Russia Western Civilization	30.95	31.95
Enterprise Information Systems	155.75	126.97

I dati sono disponibili nel file **libri-di-testo.csv** nella pagina Moodle del corso.

Soluzione. Calcoliamo la differenza dei prezzi per ciascuno dei 15 libri e riduciamo il problema ad una verifica d'ipotesi sulla differenza media dei prezzi μ_d, ovvero

$$\mathrm{H}_0 : \mu_d = 0 \text{ contro } \mathrm{H}_A : \mu_d > 0.$$

Leggiamo i dati in **R**:

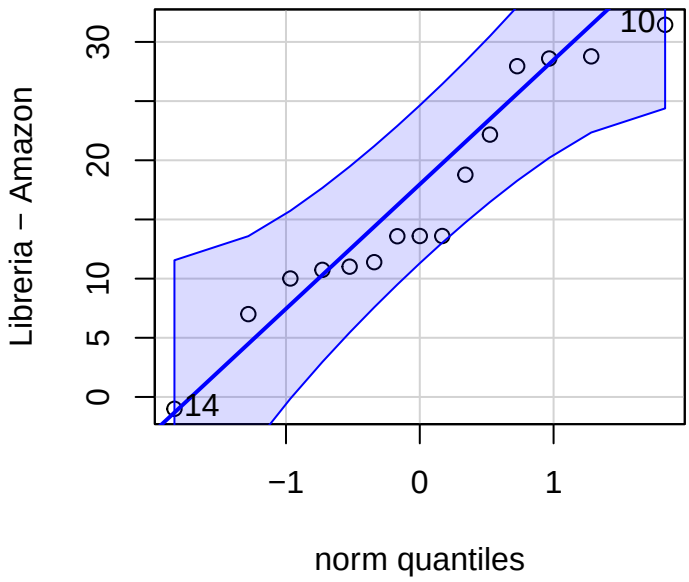
```
R> libri <- read.csv("libri-di-testo.csv")
R> head(libri)
```

```
##                            Libro Libreria Amazon
## 1    Principles of Microeconomics   120.00 101.22
## 2 Calculus: Early Transcendentals   137.50 115.33
## 3                 Exploring Wine    65.00  37.05
## 4            Manual de Gramatica    82.75  71.36
## 5               Deviant Behavior    90.00  83.00
## 6  Modern Architecture Since 1900    39.95  26.37
```

Verifichiamo la normalità delle differenze di prezzo

```
R> library(car)
R> with(libri, qqPlot(Libreria - Amazon))
```

```
## [1] 14 10
```



Il grafico quantile-quantile suggerisce che l'ipotesi di normalità non può essere rifiutata. Quindi, valutiamo il sistema d'ipotesi con un test T

$$T = \frac{\sqrt{15}\bar{D}}{S_d},$$

dove $\bar{D}$ è la media campionaria delle differenze di prezzo e S_d la deviazione standard campionaria delle differenze di prezzo. Possiamo calcolare il valore osservato della statistica T e il corrispondente p-value con la funzione **t.test**:

```
R> with(libri, t.test(Libreria - Amazon, alternative = "greater"))

##
##  One Sample t-test
##
## data:  Libreria - Amazon
## t = 6.7876, df = 14, p-value = 4.385e-06
## alternative hypothesis: true mean is greater than 0
## 95 percent confidence interval:
##  12.22633      Inf
## sample estimates:
## mean of x
##  16.51067
```

Verifichiamo il risultato:

```
R> t <- with(libri, sqrt(15) * mean(Libreria - Amazon) /
+ sd (Libreria - Amazon))
R> t
```

```
## [1] 6.787602

R> 1 - pt(t, df = 14)

## [1] 4.385292e-06
```

Il livello di significatività osservato è molto basso ad indicare un netto rifiuto dell'ipotesi nulla. Possiamo, ragionevolmente, ritenere che i prezzi dei libri di testo siano più bassi su Amazon.

Avremmo potuto ottenere lo *stesso* risultato senza calcolare le differenze dei prezzi utilizzando l'opzione **paired = TRUE**:

```
R> with(libri, t.test(Libreria, Amazon, alternative = "greater",
+ paired = TRUE))

##
##  Paired t-test
##
## data:  Libreria and Amazon
## t = 6.7876, df = 14, p-value = 4.385e-06
## alternative hypothesis: true mean difference is greater than 0
## 95 percent confidence interval:
##  12.22633      Inf
## sample estimates:
## mean difference
##        16.51067
```

10. La seguente tabella riporta i salari di sei ingegneri informatici neoassunti selezionati casualmente da due diverse aziende. I salari sono espressi in migliaia di dollari americani:

Azienda A	79	79	93	79	106	87
Azienda B	75	86	92	75	73	54

Assumendo che entrambi i campioni siano normalmente distribuiti, si risponda ai seguenti quesiti:

(a) Si calcoli un intervallo di confidenza di livello 95% per il salario medio di un ingegnere informatico neoassunto in ciascuna delle due aziende.

(b) Si valuti se i dati campionari indicano che vi sia una qualche differenza nel salario medio di un ingegnere informatico neoassunto nelle due aziende.

Soluzione

(a) La media campionaria per l'azienda A è pari a 87.16 migliaia di dollari con una deviazione standard di 10.85 migliaia di dollari. L'intervallo di confidenza di livello 95% per il salario medio dell'azienda A è

$$87.16 \pm t_{0.025}\frac{10.85}{\sqrt{6}} = 87.16 \pm 2.57\left(\frac{10.85}{\sqrt{6}}\right) = [75.78, 98.56] \text{ migliaia di dollari.}$$

La media campionaria per l'azienda B è pari a 75.83 migliaia di dollari con una deviazione standard di 13.04 migliaia di dollari. L'intervallo di confidenza di livello 95% per il salario medio dell'azienda B è

$$75.83 \pm t_{0.025}\frac{13.04}{\sqrt{6}} = 75.83 \pm 2.57\left(\frac{13.04}{\sqrt{6}}\right) = [62.14, 89.52] \text{ migliaia di dollari.}$$

Calcoli con **R**:

```
R> A <- c(79,  79,  93,  79, 106,  87)
R> mean(A)
## [1] 87.16667
R> sd(A)
## [1] 10.85204
R> qt(0.975, df = 5)
## [1] 2.570582
R> mean(A) - qt(0.975, df = 5) * sd(A) / sqrt(6)
## [1] 75.77815
R> mean(A) + qt(0.975, df = 5) * sd(A) / sqrt(6)
## [1] 98.55518
R> ##
R> B <- c(75, 86, 92, 75, 73, 54)
R> mean(B)
## [1] 75.83333
R> sd(B)
## [1] 13.04479
R> mean(B) - qt(0.975, df = 5) * sd(B) / sqrt(6)
## [1] 62.14366
R> mean(B) + qt(0.975, df = 5) * sd(B) / sqrt(6)
## [1] 89.52301
R> ## o piu' velocemente con la funzione t.test
R> t.test(A)
##
##  One Sample t-test
##
## data:  A
## t = 19.675, df = 5, p-value = 6.263e-06
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  75.77815 98.55518
## sample estimates:
## mean of x
##  87.16667
```

```
R> t.test(B)

##
##  One Sample t-test
##
## data:  B
## t = 14.24, df = 5, p-value = 3.077e-05
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  62.14366 89.52301
## sample estimates:
## mean of x
##  75.83333
```

(b) La domanda corrisponde alla verifica del sistema d'ipotesi

$$\mathrm{H}_0 : \mu_A = \mu_B \quad \text{contro} \quad \mathrm{H}_0 : \mu_A \neq \mu_B.$$

Le ipotesi possono essere valutare con la statistica test

$$T = \frac{\bar{X}_A - \bar{X}_b}{\sqrt{\frac{S_A^2}{n_A} + \frac{S_B^2}{n_B}}}$$

la cui distribuzione sotto l'ipotesi nulla può essere approssimata da una distribuzione T con gradi di libertà dati dalla formula di Sattherwaite:

```
R> t.test(A, B)

##
##  Welch Two Sample t-test
##
## data:  A and B
## t = 1.636, df = 9.6794, p-value = 0.1339
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -4.171453 26.838119
## sample estimates:
## mean of x mean of y
##  87.16667  75.83333
```

Il valore osservato della statistica T è 1.636 con un corrispondente livello di significatività osservato approssimativamente pari a 0.13. Questo p-value non permette di rifiutare l'ipotesi nulla che non vi sia una differenza significativa nei salari medi degli ingegneri informatici neoassunti dalle due aziende.