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

I Appello di Statistica per Ingegneria Energetica
2 marzo 2016

Cognome, Nome e Numero di matricola:

Problema 1. Un trifoglio ha un numero X di petali casuale con distribuzione del tipo

$$p(3) = \mathbb{P}(X = 3) = \frac{\theta}{4} \qquad p(4) = \mathbb{P}(X = 4) = \frac{\theta}{18} \qquad p(5) = \mathbb{P}(X = 5) = \frac{\theta}{36}.$$

(a) Trovare il valore di θ tale che p sia effettivamente una distribuzione di probabilità.

Sia ora θ fissato al valore trovato al punto (a).

(b) Calcolare valore atteso e varianza di X.

(c) Scrivere la funzione di ripartizione associata alla variabile X e tracciarne un grafico qualitativo.

(d) Calcolare la probabilità che un trifoglio abbia più di 3 petali e quella che ne abbia al massimo 2.

Sara abita vicino a un prato di trifogli e ha l'abitudine di raccogliere un trifoglio tutti i giorni. Calcolare la probabilità (eventualmente approssimata) che nel mese di aprile

(e) Sara non raccolga neanche un trifoglio con più di 3 petali,

(f) Sara raccolga più di 105 petali.

Risultati.

(a) Deve essere $\theta \geq 0$ e $\sum_{i=3}^{5} p(i) = 1$ da cui ricaviamo $\theta = 3$.

(b) Media e varianza sono:

$$\mathbb{E}(X) = \sum_{i=3}^{5} ip(i) = 10/3$$
$$\text{Var}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2 = 7/18.$$

(c) La funzione di ripartizione è:

$$F_X(x) = \mathbb{P}(X \leq x) = \begin{cases} 0 & \text{se } x < 3 \\ \dfrac{3}{4} & \text{se } 3 \leq x < 4 \\ \dfrac{11}{12} & \text{se } 4 \leq x < 5 \\ 1 & \text{se } x \geq 5 \end{cases}$$

(d) Le probabilità richieste sono:

$$\mathbb{P}(X > 3) = 1 - \mathbb{P}(X \leq 3) = 1 - 3/4 = 1/4$$
$$\mathbb{P}(X \leq 2) = 0$$

(e) Siano

- Y = numero di trifogli regolari su 30 $\sim B(30, 3/4)$
- W = numero di trifogli fortunati su 30 $\sim B(30, 1/4)$

Allora la probabilità che Sara non raccolga neanche un trifoglio con più di 3 petali è:

$$\mathbb{P}(Y = 30) = \mathbb{P}(W = 0) = \Big(\mathbb{P}(X = 3)\Big)^{30} = (3/4)^{30} = 0.00018 = 0.018\%$$

(f) I giorni considerati, anche se al limite, sono abbastanza numerosi ($n = 30 \geq 30$) per assumere che il numero totale T di petali raccolti in 30 giorni sia approssimativamente normale, grazie al TCL:

$$T = \sum_{j=1}^{30} X_j \simeq N\Big(10 \cdot 30/3\,,\, 7 \cdot 30/18\Big)$$

La probabilità richiesta quindi è:

$$\begin{aligned}\mathbb{P}\left(T > 105\right) &= \mathbb{P}\left(\sum_{j=1}^{30} X_j > 105\right) = 1 - \mathbb{P}\left(\sum_{j=1}^{30} X_j \leq 105\right) = 1 - \mathbb{P}\left(\sum_{j=1}^{30} X_j \leq 105.5\right) \\ &\simeq 1 - \mathbb{P}\left(Z \leq \frac{105.5 - 100}{\sqrt{\frac{7\cdot 30}{18}}}\right) = 1 - \mathbb{P}\left(Z \leq 5.5\sqrt{\frac{3}{35}}\right) \\ &= 1 - \Phi(1.61) = 1 - 0.9463 = 0.0537\end{aligned}$$

dove, come al solito, $Z \sim N(0, 1)$.

Problema 2. Franco Sottobosco e Erica Grigiobarra sono i candidati democratici per le presidenziali degli Stati Uniti per il 2016 e si devono affrontare nelle primarie. I primi stati in cui si voterà sono l'Iowa e il New Hampshire, per cui Dario Timbratore, il braccio destro di Sottobosco, è interessato alle preferenze dell'elettorato (totale, non solo quello democratico) nei due stati e, in particolare, vuole confrontare la proporzione p_I di favorevoli a Sottobosco in Iowa con la proporzione p_N in New Hampshire. Dario Timbratore commissiona quindi due sondaggi indipendenti, uno in Iowa su un campione casuale di numerosità n_I, uno in New Hampshire su un campione casuale di numerosità n_N. Siano X_I e X_N il numero di intervistati che risulteranno favorevoli a Sottobosco in Iowa e in New Hampshire rispettivamente.

(a) Specificare le leggi esatte e approssimate (per numerosità sufficientemente grande) di X_I e X_N.

(b) Introdurre, in funzione di X_I e X_N, degli opportuni stimatori non distorti $\widehat{p}_I$ e $\widehat{p}_N$ di p_I e p_N.

(c) Impostare un opportuno test di ipotesi ad un generico livello di significatività α per aiutare il sig. Timbratore a stabilire se la proporzione di favorevoli a Sottobosco in Iowa sia maggiore rispetto a quella in New Hampshire. Specificare:

- ipotesi nulla e ipotesi alternativa;
- regione critica di livello α;
- condizioni di applicabilità.

(d) In Iowa, su 1000 intervistati, 530 risultano favorevoli a Sottobosco, mentre in New Hampshire i favorevoli a Sottobosco sono 576 su 1200 intervistati.

- Calcolare le stime puntuali delle proporzioni di favorevoli a Sottobosco in Iowa e New Hampshire;
- controllare se le condizioni di applicabilità del test al punto (c) sono verificate per i dati raccolti;
- calcolare il p-value dei dati per il test al punto (c);
- cosa può concludere il sig. Timbratore a un livello di significatività del 5%?
- Si tratta di una conclusione debole o forte?

(e) Costruire un intervallo di confidenza al 95% per la differenza tra le due proporzioni.

Risultati.

(a) Le variabili aleatorie oggetto dell'analisi sono:

- X_I = n° di favorevoli in Iowa $\sim B(n_I, p_I) \simeq N(n_I p_I, n_I p_I(1-p_I))$
- X_N = n° di favorevoli in New Hampshire $\sim B(n_N, p_N) \simeq N(n_N p_N, n_N p_N(1-p_N))$

(b) Gli stimatori puntuali non distorti sono:

- $\widehat{p}_I = X_I/n_I$
- $\widehat{p}_N = X_N/n_N$

(c)

- Ipotesi nulla e ipotesi alternativa:

$$H_0 : p_I \leq p_N \qquad\qquad H_1 : p_I > p_N$$

- La regione critica di livello α è:

$$RC_\alpha = \left\{ \hat{p}_I > \hat{p}_N + \sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_I} + \frac{1}{n_N}\right)}\, z_\alpha \right\} = \left\{ z_0 = \frac{\hat{p}_I - \hat{p}_N}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_I} + \frac{1}{n_N}\right)}} > z_\alpha \right\},$$

dove

$$\hat{p} = \frac{X_I + X_N}{n_I + n_N}.$$

- Condizioni di applicabilità del test: campioni casuali provenienti da popolazioni indipendenti e campioni numerosi ($n_I\widehat{p}_I$, $n_I(1-\widehat{p}_I)$, $n_N\widehat{p}_N$, $n_N(1-\widehat{p}_N) > 5$).

(d)

- Le stime delle proporzioni di favorevoli a Sottobosco in Iowa e New Hampshire sono:

$$\widehat{p}_I = \frac{530}{1000} = 0.53 \qquad \widehat{p}_N = \frac{576}{1200} = 0.48.$$

- Dato che $n_I\widehat{p}_I$, $n_I(1-\widehat{p}_I)$, $n_N\widehat{p}_N$ e $n_N(1-\widehat{p}_N)$ sono tutti maggiori di 5 possiamo usare il TCL; inoltre X_I e X_N sono indipendenti e quindi le condizioni di applicabilità del test sono verificate.
- Il p-value dei dati è:

$$\text{p-value} = \mathbb{P}(Z > z_0) = \mathbb{P}(Z > 2.34) = 0.0096.$$

- Dato che 0.05 > p-value, rifiutiamo H_0 al 5% e quindi il sig. Timbratore può concludere che in Iowa la percentuale di favorevoli a Sottobosco è maggiore che in New Hampshire:

$$p_I > p_N.$$

- Conclusione forte.

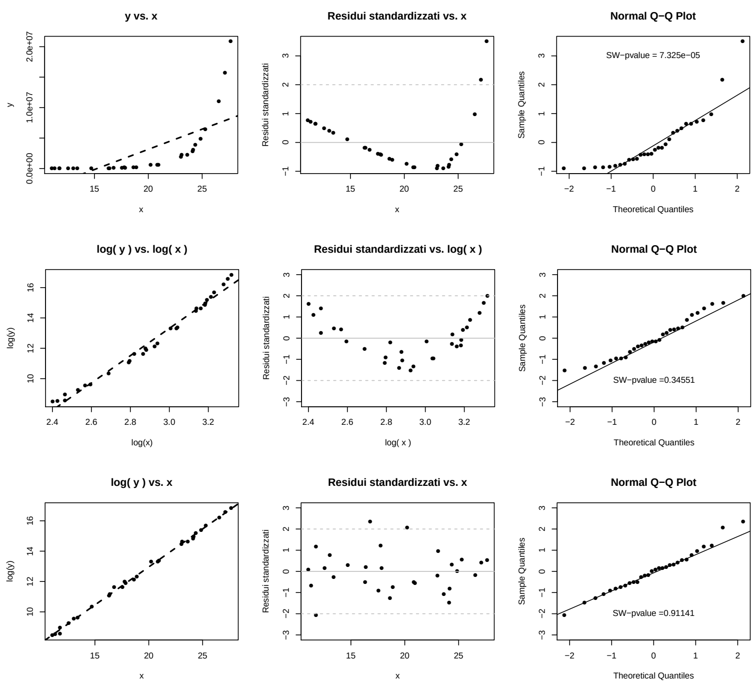
(e) Grazie alla numerosità dei campioni ($n_I\widehat{p}_I$, $n_I(1-\widehat{p}_I)$, $n_N\widehat{p}_N$, $n_N(1-\widehat{p}_N) > 5$), l'intervallo di confidenza richiesto è:

$$\begin{aligned} IC_{95\%}(p_I - p_N) &= \hat{p}_I - \hat{p}_N \pm z_{0.025}\sqrt{\frac{\hat{p}_I(1-\hat{p}_I)}{n_I} + \frac{\hat{p}_N(1-\hat{p}_N)}{n_N}} \\ &= 0.05 \pm 1.96 \cdot 0.02 \approx 0.05 \pm 0.04 = (0.01\,,\, 0.09). \end{aligned}$$

Problema 3. La gente della Contea è conosciuta, almeno nelle immediate vicinanze, come gente semplice che ama la tranquillità, il buon cibo, la buona birra e l'erba pipa. Ultimamente però alcuni abitanti hanno iniziato a discutere animatamente su quale sia la giusta relazione fra il numero di boccali di birra X bevuti da un mezzuomo nell'arco di 24 ore e il corrispondente apporto calorico Y (espresso in opportune unità di misura). Oggetto della discussione e della contesa sono tre possibili modelli empirici gaussiani di regressione lineare. Per dirimere la questione sono stati selezionati casualmente 30 mezzuomini e, per ciascuno di loro, ieri sono stati raccolti i valori di X e Y, che hanno dato medie campionarie

$$\overline{x}_{30} = 19.19, \qquad \overline{y}_{30} = 2607583,$$

e che, elaborati secondo i tre modelli, hanno fornito i seguenti risultati:




```
Residuals:
     Min       1Q   Median       3Q      Max
-3261386 -2559682 -1159594  1714646 12657095

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -10116267    2550238  -3.967 0.000459 ***
x              663088     128376   5.165 1.77e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3615000 on 28 degrees of freedom
Multiple R-squared:  0.4879,    Adjusted R-squared:  0.4696
F-statistic: 26.68 on 1 and 28 DF,  p-value: 1.766e-05
```

```
Residuals:
     Min       1Q   Median       3Q      Max
-0.51698 -0.28632 -0.05364  0.16495  0.67054

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -13.6631     0.6335  -21.57   <2e-16 ***
log(x)        8.9917     0.2162   41.58   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3369 on 28 degrees of freedom
Multiple R-squared:  0.9841,    Adjusted R-squared:  0.9835
F-statistic:  1729 on 1 and 28 DF,  p-value: < 2.2e-16
```

```
Residuals:
      Min        1Q    Median        3Q       Max
-0.250235 -0.078283  0.006104  0.061547  0.285783

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 2.937552   0.085358   34.41   <2e-16 ***
x           0.501084   0.004297  116.62   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.121 on 28 degrees of freedom
Multiple R-squared:  0.9979,    Adjusted R-squared:  0.9979
F-statistic: 1.36e+04 on 1 and 28 DF,  p-value: < 2.2e-16
```

Data la loro goffaggine con la matematica, i mezzuomini chiedono il vostro aiuto.

(a) Scrivere la relazione tra le variabili X e Y ipotizzata dai tre modelli.
   In particolare, esplicitare Y in funzione di X in tutti e tre i modelli.

(b) Ordinare i modelli in ordine decrescente (dal migliore al peggiore) in base a chi meglio spiega la variabilità della risposta.

(c) Indicare quali modelli hanno residui omoschedastici.

(d) Indicare per quali modelli appare soddisfatta l'ipotesi gaussiana.

(e) Per ogni modello cerchiare eventuali outlier nel grafico di dispersione dei residui.

(f) Quale modello risulta quindi migliore alla luce dei dati raccolti?

Sam ha perso una scommessa e domani potrà bere solo 12 birre.

(g) Fornire una previsione puntuale del suo apporto calorico.

(h) Fornire anche una previsione intervallare al 90%.

Risultati.

(a) Indicando con $\epsilon \sim N(0, \sigma^2)$ l'errore gaussiano presente in ciascun modello, possiamo scrivere

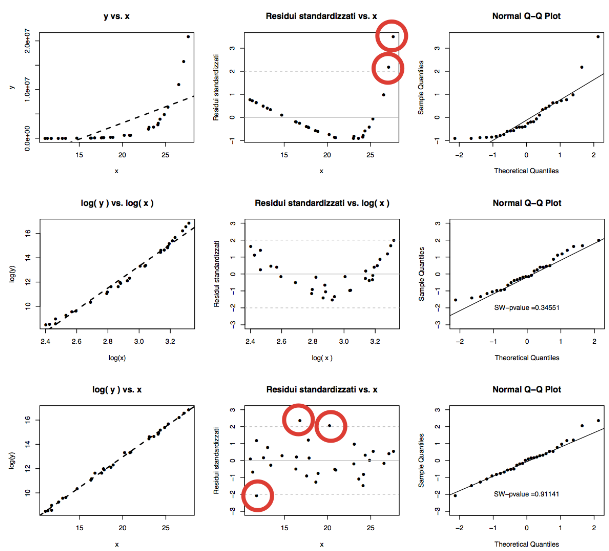
$$
\begin{aligned}
&\text{modello 1:} \quad && Y = \beta_0 + \beta_1 X + \epsilon, \\
&\text{modello 2:} \quad && \log(Y) = \beta_0 + \beta_1 \log(X) + \epsilon \implies Y = \mathrm{e}^{\beta_0+\epsilon} X^{\beta_1} \\
&\text{modello 3:} \quad && \log(Y) = \beta_0 + \beta_1 X + \epsilon \implies Y = \exp\Big\{\beta_0 + \beta_1 X + \epsilon\Big\}
\end{aligned}
$$

(b) Modello 3 ($R^2 = 0.998$), modello 2 ($R^2 = 0.984$) e modello 3 ($R^2 = 0.488$).

(c) Solo il modello 3 presenta residui omoschedastici.

(d) Solo per il modello 3 appare soddisfatta l'ipotesi gaussia: residui omoschedastici, con buon NPP, e alto p-value (0.91) nel test di SW.

(e) Outlier:



(f) Modello 3.

(g) Utilizzando il modello 3:

$$\widehat{\log(Y)} = \widehat{\beta_0} + \widehat{\beta_1} x_0 = 2.937552 + 0.501084 \cdot 12 = 8.950562 \implies \widehat{Y} = \exp(8.950562) = 7712$$

(h) Utilizziamo sempre il modello 3. Invertendo la formula di $\mathrm{se}(\widehat{\beta_1})$ si può ricavare S_{xx}:

$$\mathrm{se}(\widehat{\beta_1}) = \sqrt{\frac{\hat{\sigma}^2}{S_{xx}}} \implies S_{xx} = \frac{\hat{\sigma}^2}{\mathrm{se}(\widehat{\beta_1})^2} = \frac{0.121^2}{0.004297^2} = 792.9395$$

L'intervallo di previsione al 90% per $\log(Y)$ è dato da

$$
\begin{aligned}
\widehat{\beta_0} + \widehat{\beta_1} x_0 \pm \hat{\sigma} \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}}\, t_{\alpha/2}(n-2) &= 8.950562 \pm 0.121 \sqrt{1 + \frac{1}{30} + \frac{(12 - 19.19)^2}{792.9395}}\, t_{0.05}(29) \\
&= 8.950562 \pm 0.121 \cdot \sqrt{1.0985} \cdot 1.69 = 8.9506 \pm 0.2154 = (8.7352\,,\, 9.1660),
\end{aligned}
$$

per cui l'intervallo di previsione al 90% per Y è dato da

$$(\mathrm{e}^{8.7352}\,,\, \mathrm{e}^{9.1660}) = (6217\,,\, 9567).$$