Politecnico di Milano - Scuola di Ingegneria Industriale

II Prova in Itinere di Statistica per Ingegneria Energetica
5 luglio 2012

Cognome, Nome e Numero di matricola:

Problema 1. Andrea ha sempre desiderato essere un campione di pallacanestro e da poco gioca nella squadra della sua città. In realtà, più che giocare sta in panchina, dato l'alto livello dei suoi compagni. Spesso capita addirittura che questi lo prendano in giro, sostenendo che non realizzi neanche la metà dei tiri liberi che prova. Un giorno Andrea, per mettere fine a queste voci, effettua 50 tiri liberi, realizzandone 30, affermando poi: "avete visto?"

I compagni di Andrea hanno però seguito con profitto un corso di Statistica e sanno che tale risultato potrebbe essere semplicemente frutto di 50 tiri fortunati. Sanno quindi che, dandogli ragione, potrebbero cessare erroneamente di prenderlo in giro e sono disposti a commettere tale errore solo con una probabilità del 5%.

(a) Andrea è quindi riscito a convincere i suoi compagni che avevano torto? Si risponda eseguendo un opportuno test di ipotesi, esplicitando chiaramente: il parametro incognito, le ipotesi statistiche, la regione critica per il livello di significatività richiesto, la conclusione.

(b) Qual è il numero minimo di tiri da realizzare, su 50 effettuati, per mettere fine a queste voci.

(c) Se in verità Andrea è in grado di realizzare il 70% dei tiri liberi che prova, con quale probabilità può mettere fine a queste voci con 50 tiri a disposizione.

Risultati.

(a) Siano p la proporzione incognita di tiri liberi che Andrea è in grado di fare e $p_0 = 0.5$. Si deve eseguire un test per la verifica delle ipotesi $H_0 : p < p_0$ vs $H_1 : p \geq p_0$, di livello $\alpha = 0.05$. La regione critica è quindi

$$\frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} > z_\alpha,$$

purché il campione sia sufficientemente numeroso, ovvero purché il numero atteso di successi e di insuccessi per $p = p_0$ sia almeno 5. Nel nostro caso i successi e gli insuccessi attesi per $p = 0.5$ sono 25 e la statistica test vale

$$z_0 = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{0.6 - 0.5}{\sqrt{\frac{0.5(1-0.5)}{50}}} = 1.414$$

da confrontare con il punto percentuale $z_\alpha = 1.645$: non vi è evidenza statistica per rifiutare H_0, Andrea continuerà ad essere preso in giro.

(b) Si deve cercare il minimo numero naturale k tale che

$$\frac{k}{50} = \hat{p} > p_0 + \sqrt{\frac{p_0(1-p_0)}{50}}\, 1.645 \qquad \Longleftrightarrow \qquad k > 30.8$$

per cui il numero minimo è $k = 31$.

(c) Si deve calcolare la potenza del test per $p_1 = 0.7$. Detto X il numero di tiri realizzati su 50, si ha $X \sim B(50, 0.7)$ per cui,

$$\text{potenza} = P\left(X > 30.8\right) = P\left(X > 30.5\right) \simeq 1 - \Phi\left(\frac{30.5 - 35}{\sqrt{10.5}}\right) \simeq \Phi(1.39) = 0.9177.$$

Oppure, trascurando la correzione di continuità, si trova:

$$P_{p=p_1}(Z_0 > z_\alpha) = P_{p=p_1}\left(\hat{p} > z_\alpha\sqrt{\frac{p_0(1-p_0)}{n}} + p_0\right)$$

$$= P_{p=p_1}\left(\frac{\hat{p}-p_1}{\sqrt{\frac{p_1(1-p_1)}{n}}} > z_\alpha\sqrt{\frac{p_0(1-p_0)}{p_1(1-p_1)}} + \frac{p_0-p_1}{\sqrt{\frac{p_1(1-p_1)}{n}}}\right)$$

$$= 1 - \phi\left(z_\alpha\sqrt{\frac{p_0(1-p_0)}{p_1(1-p_1)}} + \frac{p_0-p_1}{\sqrt{\frac{p_1(1-p_1)}{n}}}\right),$$

con $p_0 = 0.5$, $p_1 = 0.7$, $n = 50$, $\alpha = 0.05$ da cui:

$$\text{potenza} = 1 - \phi\left(1.645\sqrt{\frac{0.25}{0.21}} + \frac{-0.2}{\sqrt{\frac{0.21}{50}}}\right) =$$

$$1 - \phi(1.7947 - 3.086) = 1 - \phi(-1.2914) = \phi(1.2914) = 0.9017$$

Problema 2. Il Dott. Piccolini, famoso naturalista, si sta recentemente occupando delle relazioni zanzare e pipistrelli. Per questo studio, avrebbe bisogno in particolare di conoscere il rapporto tra il numero di pipistrelli e il numero di zanzare in uno specifico quartiere di Roma alla mezzanotte tra il 23 e il 24 giugno 2012. Purtroppo, il numero di pipistrelli p e il numero di zanzare z sono incogniti ed è quindi necessario stimarli per ottenere il rapporto $r = \frac{z}{p}$. Perciò il Dott. Piccolini decide di inviare 11 giovani ricercatori a misurare la presenza di zanzare e 11 giovani ricercatori (diversi dai precedenti) a misurare il numero di pipistrelli in quel quartiere e a quell'ora. Ciascun ricercatore è dotato di uno strumento (coperto da segreto) che permette di ottenere una misura del numero di rappresentanti di una specie nella zona considerata. Supponiamo che che le misure $P_1, \ldots, P_{11}$ del numero di pipistrelli siano indipendenti, con errore sistematico nullo e varianza σ_P^2 e che le misure $Z_1, \ldots, Z_{11}$ del numero di zanzare siano indipendenti, con errore sistematico nullo e varianza σ_Z^2.

(a) Proporre stimatori puntuali per p, z e r.

(b) Verificare se tali stimatori sono, almeno approssimativamente, non distorti.

(c) Calcolare l'errore quadratico medio, almeno approssimato, di questi stimatori.

Le misure raccolte dai ricercatori forniscono come medie campionarie rispettivamente $\bar{p}_{11} = 10$ e $\bar{z}_{11} = 1000$ e come varianze campionarie $s_P^2 = 1.10$ e $s_Z^2 = 132.00$.

(d) Fornire delle stime puntuali per p, z e r.

(e) Stimare l'errore quadratico medio degli stimatori utilizzati.

Risultati.

(a) $\widehat{p} = \overline{P}_{11}$, $\widehat{z} = \overline{Z}_{11}$, $\widehat{r} = \frac{\overline{Z}_{11}}{\overline{P}_{11}}$.

(b) Dato che l'errore sistematico è ipotizzato nullo, $\mathbb{E}[\widehat{p}] = \mathbb{E}[\overline{P}_{11}] = p$ e $\mathbb{E}[\widehat{z}] = \mathbb{E}[\overline{Z}_{11}] = z$.
La media dello stimatore $\widehat{r}$, approssimata con il metodo delta, è : $\mathbb{E}[\widehat{r}] \approx \frac{\mathbb{E}[\overline{Z}_{11}]}{\mathbb{E}[\overline{P}_{11}]} = \frac{z}{p} = r$.

(c) $\mathrm{MSE}(\widehat{p}) = \mathrm{Var}(\widehat{p}) = \frac{\sigma_P^2}{11}$ e $\mathrm{MSE}(\widehat{z}) = \mathrm{Var}(\widehat{z}) = \frac{\sigma_Z^2}{11}$. Se $h(x,y) = \frac{x}{y}$,
$\mathrm{MSE}(\widehat{r}) \approx \mathrm{Var}(\widehat{r}) \approx (\frac{\partial h}{\partial x})^2|_{x=z,y=p}\mathrm{Var}(\overline{Z}_{11}) + (\frac{\partial h}{\partial y})^2|_{x=z,y=p}\mathrm{Var}(\overline{P}_{11}) = \frac{1}{p^2}\frac{\sigma_Z^2}{11} + \frac{z^2}{p^4}\frac{\sigma_P^2}{11}$.

(d) $\widehat{p} = \overline{p}_{11} = 10$, $\widehat{z} = \overline{z}_{11} = 1000$, $\widehat{r} = \frac{\overline{z}_{11}}{\overline{p}_{11}} = 100$.

(e) $\mathrm{MSE}(\widehat{p}) \approx \frac{s_P^2}{11} = 0.1$, $\mathrm{MSE}(\widehat{z}) \approx \frac{s_Z^2}{11} = 12$, $\mathrm{MSE}(\widehat{r}) \approx \frac{1}{\widehat{p}^2}\frac{s_Z^2}{11} + \frac{\widehat{z}^2}{\widehat{p}^4}\frac{s_P^2}{11} = 10.12$

Problema 3. La IceStats SpA è una nota catena di gelaterie milanesi, molto famose per la particolarità dei gusti serviti. Il dott. Fresco, proprietario della IceStats, è da sempre interessato a valutare il rapporto che intercorre tra la quantità di gelato venduto, e alcune grandezze significative. Per questo, fa raccogliere i dati relativi alla vendita di gelati, temperatura esterna e umidità relativa dell'aria nelle due stagioni estive 2010 e 2011 (da maggio a settembre, periodo di apertura della IceStats). In particolare, per ognuno dei 296 giorni di apertura, ha a disposizione il valore di totale di gelato venduto nelle sue gelaterie G (in Kg.), la temperatura media giornaliera T (in °C) e l'umidità relativa dell'aria U (percentuale) a Milano, e vuole valutare la dipendenza di G dalle altre due variabili. Medie campionarie e varianze campionarie dei dati raccolti sono:

$$\overline{g} = 182.1922,\ s_g^2 = 1390.459, \qquad \overline{t} = 17.61963,\ s_t^2 = 14.1843, \qquad \overline{u} = 62.88284,\ s_u^2 = 88.23529$$

In Figg. 1 e 2 vengono proposti gli output di R del modello di regressione, il grafico dei residui e i p-value del test di Shapiro-Wilk per i residui per ciascuno dei seguenti modelli:

$$\text{Modello 1: } G_i = \beta_0 + \beta_1 T_i + \beta_2 U_i + \epsilon_i$$

$$\text{Modello 2: } G_i = \beta_0 + \beta_1 T_i + \epsilon_i$$

con $\epsilon_i \sim N(0, \sigma^2)$, per $i = 1, ..., 296$.

(a) Si commenti la bontà dei modelli proposti e si scelga di conseguenza il migliore per descrivere il problema in esame.

(b) Si scriva l'equazione di regressione stimata per il modello prescelto.

(c) La stazione meteorologica di Milano Linate stima che per domani la temperatura media a Milano sarà di 22 gradi e l'umidità relativa del 85%. Fornire una previsione intervallare al 90% della quantità di gelato che sarà venduta, in totale, nella giornata di domani.

Al momento dell'apertura di 15 anni fa, il dott. Fresco aveva già fatto un'analisi simile, ottenendo la seguente stima per il coefficiente relativo alla temperatura: $\hat{\beta}_1^{old} = 9$.

(d) È possibile affermare che il coefficiente di dipendenza lineare sia cambiato rispetto a quello di 15 anni fa? Effettuare un opportuno test al 5%.

Soluzione:

(a) Il Modello 1 ha un R2 adjusted molto elevato, i residui non presentano particolari trend. Considerando il p-value del test di Shapiro-Wilks, l'ipotesi della normalità dei residui non è rifiutata a tutti i livelli usuali. Pertanto è possibile considerare i test di significatività proposti nell'output di R. Il modello è globalmente significativo (p-value: $< 2.2e-16$), ma il coefficiente β_2 risulta non significativamente diverso da 0 (p-value: 0.4525). Per questo motivo sarebbe opportuno eliminare il predittore U dal modello.

Il Modello 2, ottenuto proprio eliminando U, presenta le stesse buone caratteristiche del Modello 1, ma in questo caso tutti i predittori risultano significativi. Inoltre R2 adjusted è leggermente aumentato. É un modello più semplice del Modello 1, avendo un regressore in meno, ma ha le stesse performances, se non leggermente superiori. Per questi motivi è opportuno scegliere il Modello 2.

(b) L'equazione stimata per il modello 2 è la seguente:

$$\hat{G} = 14.1758 + 9.5358 \times T$$

(c) La previsione puntuale per la quantità di gelato venduta è $\hat{G}_0 = \hat{G}_{|T=22} = 14.1758 + 9.5358 \times 22 = 223.9634Kg$. L'intervallo di previsione cercato è: $\hat{G}_0 \pm t_{0.025,n-2}\hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(t_0-\overline{t})^2}{S_{tt}}}$, con $S_{tt} = s_t^2(n-1)$, da cui l'intervallo cercato: $[207.315, 240.6118]$.

(d) Il test di ipotesi è il seguente: $H0 : \beta_1 = 9$ vs. $H1 : \beta_1 \neq 9$. Quindi, eseguendo un t-test sul coefficiente β_1 si ottiene: $T0 = (\hat{\beta}_1 - 9)/se(\hat{\beta}_1) = 3.447876$, $t_{0.025,n-2} = 1.960$, quindi al 5% c'è evidenza per rifiutare $H0$.

```
Call:
lm(formula = G ~ T + U)

Residuals:
    Min      1Q  Median      3Q     Max
-28.161  -6.762  -0.285   6.986  31.820

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 10.14154    6.05040   1.676   0.0948 .
T            9.58378    0.16809  57.017   <2e-16 ***
U            0.05070    0.06739   0.752   0.4525
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10.06 on 293 degrees of freedom
Multiple R-squared: 0.9277,     Adjusted R-squared: 0.9272
F-statistic:  1881 on 2 and 293 DF,  p-value: < 2.2e-16

> res1 <- mod1$residuals
> shapiro.test(res1)

        Shapiro-Wilk normality test

data:  res1
W = 0.9973, p-value = 0.9054
```

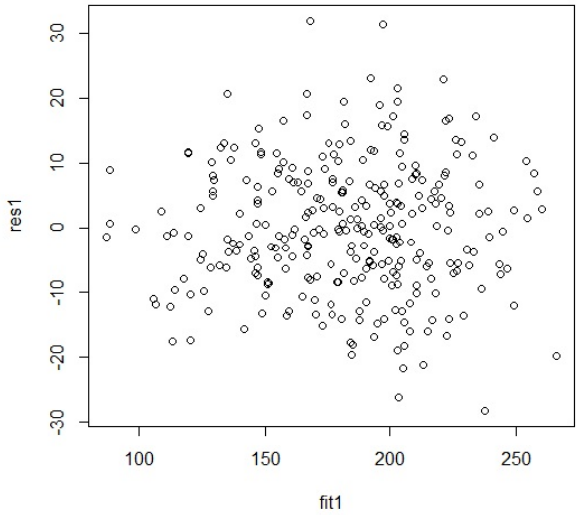

Figura 1: Output dell'analisi per il modello 1

```
Call:
lm(formula = G ~ T)

Residuals:
    Min      1Q  Median      3Q     Max
-27.481  -6.295  -0.308   6.959  31.983

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  14.1758     2.7993   5.064 7.25e-07 ***
T             9.5358     0.1554  61.373  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10.05 on 294 degrees of freedom
Multiple R-squared: 0.9276,     Adjusted R-squared: 0.9274
F-statistic:  3767 on 1 and 294 DF,  p-value: < 2.2e-16

> res2 <- mod2$residuals
> shapiro.test(res2)

        Shapiro-Wilk normality test

data:  res2
W = 0.9971, p-value = 0.8808
```

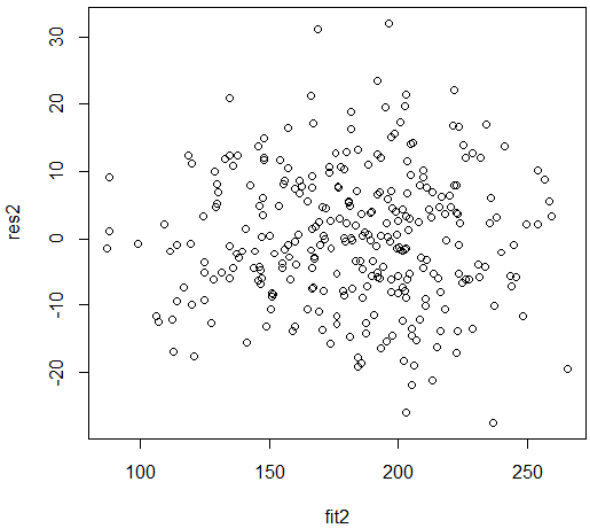

Figura 2: Output dell'analisi per il modello 2