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

I Appello di Statistica per Ingegneria Fisica
28 agosto 2017

Cognome, Nome, Matricola:

Problema 1. Un biglietto di lotteria istantanea costa 2 euro e assicura una vincità V così distribuita:

k	1	2	3	4	5
$p_V(k)$	0.4	0.4	0.1	0.05	0.05

1. Trovare media e varianza della vincita V.
2. Trovare media e varianza del guadagno G.

Si consideri il guadagno totale X_n di n biglietti indipendenti.

3. Trovare media e varianza del guadagno totale X_n.
4. Trovare, in funzione di n, la probabilità di perdere più di 15 euro.
   È possibile dare solo una risposta approssimata per n grande.
5. Calcolare il limite di tale probabilità per $n \to \infty$.
6. Per quali n tale probabilità supera il 75%?

Risultati.

1. $\mathbb{E}[V] = 1.95, \qquad \mathrm{Var}(V) = 1.1475.$
2. $\mathbb{E}[G] = \mathbb{E}[V-2] = \mathbb{E}[V] - 2 = -0.05, \qquad \mathrm{Var}(G) = \mathrm{Var}(V-2) = \mathrm{Var}(V) = 1.1475.$
3. Se G_k è il guadagno del biglietto k, allora $X_n = \sum_{k=1}^{n} G_k$ per cui

   $\mathbb{E}[X_n] = -0.05\,n, \qquad \mathrm{Var}(X_n) = 1.1475\,n.$
4. Per n grande vale il TCL per cui

   $\mathbb{P}(X_n < -15) = \mathbb{P}(X_n \leq -15.5) \simeq \Phi\left(\dfrac{-15.5 + 0.05\,n}{\sqrt{1.1475\,n}}\right).$
5. $\mathbb{P}(X_n < -15) \to 1$ per $n \to \infty$.
6. Supponendo n grande abbiamo

$$\Phi\left(\frac{-15.5 + 0.05\,n}{\sqrt{1.1475\,n}}\right) \geq 0.75 \iff \frac{-15.5 + 0.05\,n}{\sqrt{1.1475\,n}} \geq z_{0.25} = 0.674 \iff n \geq 690$$

   che è coerente con l'ipotesi n grande.

Problema 2. Il Professor Mosk Han vuole provare che la proporzione di newtype nella popolazione di Side 7 è superiore alla proporzione di newtype nella popolazione di Side 3. Non potendo ricorrere ad un censimento deve accontentarsi di una indagine campionaria e delle relative conclusioni inferenziali.

1. Impostate un opportuno test statistico per provare che la proporzione di newtype su Side 7 è superiore alla proporzione di newtype su Side 3. Specificate in particolare:
    - le distribuzioni delle popolazioni di interesse e i rispettivi parametri incogniti su cui inferire,
    - ipotesi nulla e ipotesi alternativa del test,
    - regione critica di livello α per decidere sulla base di due campioni casuali, uno per popolazione, di numerosità (elevata) n_3 ed n_7 rispettivamente.

Il Professor Mosk Han riesce a esaminare un campione casuale di $n_3 = 25$ abitanti di Side 3, trovando 8 newtype, ed un campione casuale di $n_7 = 54$ abitanti di Side 7, trovando 19 newtype.

2. Quanto valgono le proporzioni di newtype nei due campioni?
3. Quanto vale il p-value dei dati raccolti?
4. Cosa può concludere il Professor Mosk Han? La conclusione è forte o debole?

Risultati.

1.
   - Popolazioni Bernoulliane (successo = newtype) di parametri rispettivamente p_3 e p_7, dove p_k è la proporzione di newtype nella popolazione di Side k
   - $H_0 : p_7 \leq p_3 \qquad \text{vs} \qquad H_1 : p_7 > p_3$
   - Indicando con $\widehat{p}_k$ la proporzione di newtype nel campione di Side k, posto $\widehat{p} = \dfrac{n_3\,\widehat{p}_3 + n_7\,\widehat{p}_7}{n_3 + n_7}$,

     $$R_\alpha : \widehat{p}_7 > \widehat{p}_3 + \sqrt{\widehat{p}(1-\widehat{p})\left(\frac{1}{n_3} + \frac{1}{n_7}\right)}\, z_\alpha$$

2. $\widehat{p}_7 = 0.3518$ mentre $\widehat{p}_3 = 0.32$

3. Per i dati raccolti

   $$z_\alpha = \frac{\widehat{p}_7 - \widehat{p}_3}{\sqrt{\widehat{p}(1-\widehat{p})\left(\frac{1}{n_3} + \frac{1}{n_7}\right)}} = 0.28$$

   dà

   $$\text{p-value} = 1 - \Phi(0.28) = 0.39$$

4. Nonostante $\widehat{p}_7 > \widehat{p}_3$, il p-value è alto e il Professor Mosk Han <u>non</u> può rifiutare H_0 agli usuali livelli di significatività. Ottiene quindi la conclusione debole: $p_7 \leq p_3$.

Problema 3. La società Firebolt s.r.l sta sperimentando un nuovo tipo di saldatore laser applicato alla saldatura a sovrapposizione. In particolare vuole studiare la relazione tra il rapporto di forma H (cioè il rapporto fra profondità e larghezza del cordone saldato) e alcuni parametri di processo: la potenza dell'impulso p (in kW), la durata dell'impulso t (in ms) e il diametro dello spot d (in mm). Vengono considerati due possibili modelli empirici gaussiani di regressione lineare,

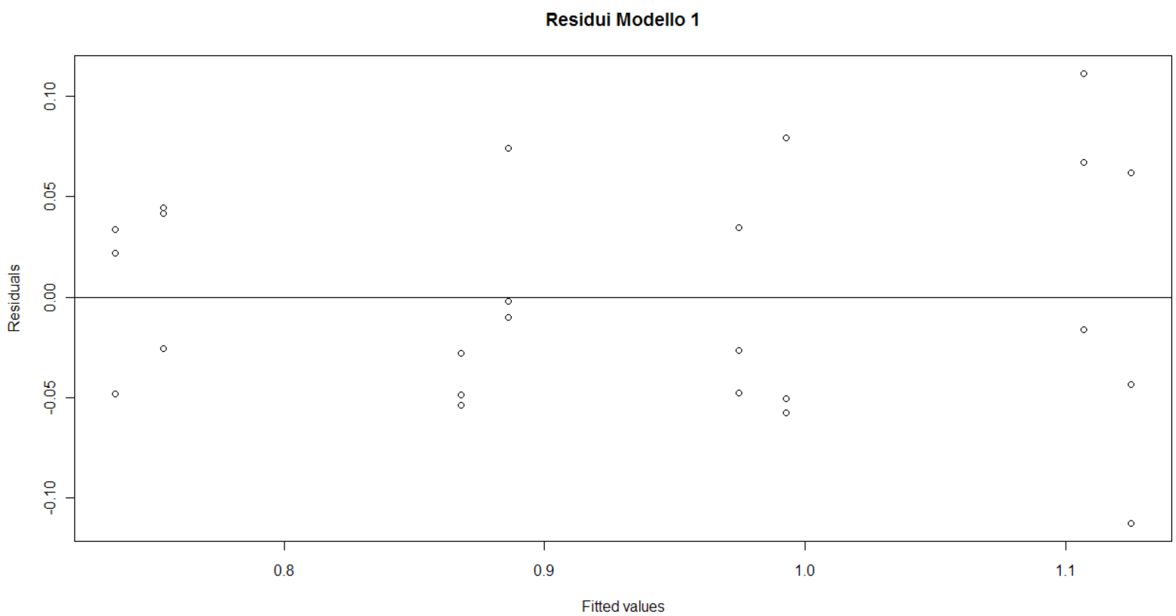
- Modello 1: H su p, t e d,
- Modello 2: H su p e d.

1. Si scriva la relazione ipotizzata dai due modelli fra il responso H e i corrispondenti predittori.

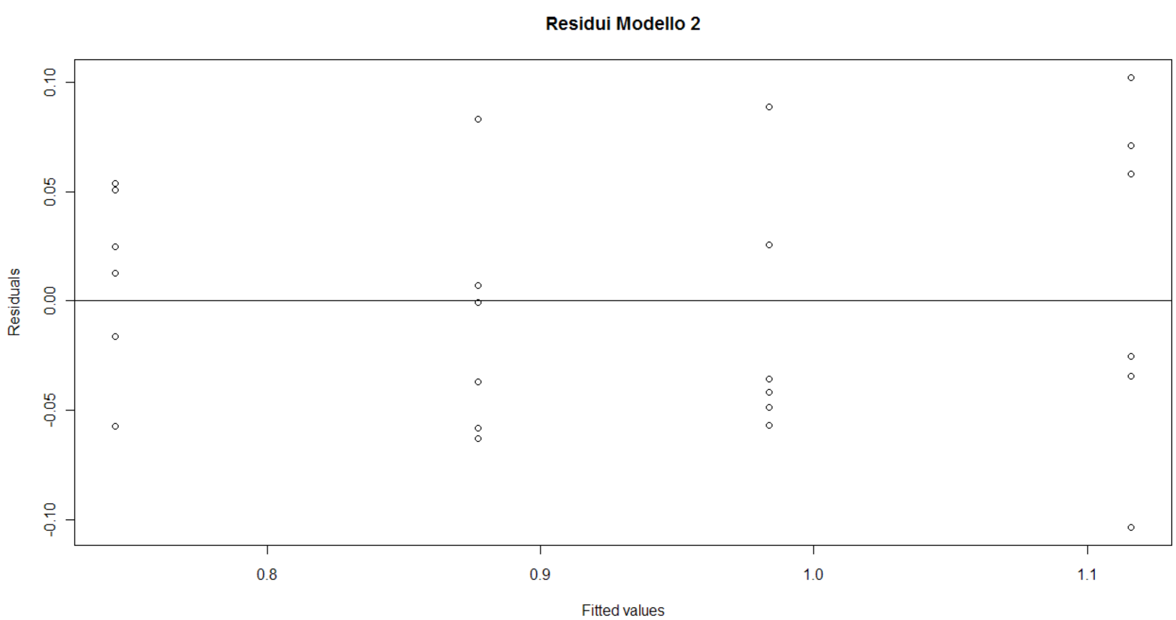
I risultati di 24 prove di laboratorio vengono quindi elaborati sulla base dei due modelli di regressione, fornendo i dati di sintesi, i p-value di Shapiro-Wilk dei residui e il diagramma di dispersione dei residui sui responsi stimati riportati di seguito.

2. Si commenti l'adeguatezza dei modelli proposti in relazione ai dati raccolti.
3. Si trovi una stima puntuale per il rapporto di forma medio nel caso $p = 1.3$, $t = 10$ e $d = 0.4$
4. Si trovi una stima puntuale per la variazione media del rapporto di forma se, a parità degli altri predittori, d aumenta di 0.1.
5. Si trovi una stima intervallare al 90% per il rapporto di forma medio nel caso $p = 0$, $t = 0$ e $d = 0$.
6. Sapendo che per il Modello 2 il residuo minimo vale -0.103125, si trovino le coordinate del punto corrispondente nel Normal Probability Plot dei residui standardizzati.

```
> summary(Model1)

Call:
lm(formula = H ~ P + T + D)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.11225 -0.04760 -0.01300  0.04233  0.11100 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.555726   0.079496   6.991 8.77e-07 ***
P            0.478500   0.048853   9.795 4.48e-09 ***
T           -0.002607   0.003489  -0.747    0.464    
D           -0.441944   0.081421  -5.428 2.59e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.05983 on 20 degrees of freedom
Multiple R-squared: 0.863,	Adjusted R-squared: 0.8424 
F-statistic: 41.99 on 3 and 20 DF,  p-value: 8.08e-09

> shapiro.test(Model1$residuals)

	Shapiro-Wilk normality test

data:  Model1$residuals
W = 0.9553, p-value = 0.3506
```



```
> summary(Model2)

Call:
lm(formula = H ~ P + D)

Residuals:
      Min        1Q    Median        3Q       Max 
-0.103125 -0.043292 -0.008583  0.051458  0.101875 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.53096    0.07149   7.427 2.65e-07 ***
P            0.47850    0.04834   9.899 2.31e-09 ***
D           -0.44194    0.08056  -5.486 1.92e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.0592 on 21 degrees of freedom
Multiple R-squared: 0.8591,	Adjusted R-squared: 0.8457 
F-statistic: 64.05 on 2 and 21 DF,  p-value: 1.153e-09

> shapiro.test(Model2$residuals)

	Shapiro-Wilk normality test

data:  Model2$residuals 
W = 0.9519, p-value = 0.2975
```



Risultati.

1.

$$\begin{aligned} &\text{Modello1}: \quad H = \beta_0 + \beta_1 p + \beta_2 t + \beta_3 d + \epsilon, \qquad && \epsilon \sim N(0, \sigma^2), \\ &\text{Modello2}: \quad H = \beta_0 + \beta_1 p + \beta_2 d + \epsilon, && \epsilon \sim N(0, \sigma^2). \end{aligned}$$

2. Il Modello 1 ha un R^2_{adj} abbastanza elevato, i residui non presentano tendenze particolari. 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 8.08×10^{-09}), ma il coefficiente β_2 risulta non significativamente diverso da 0 (p-value 0.464). Per questo motivo sarebbe opportuno eliminare il predittore t dal modello.

   Il Modello 2, ottenuto proprio eliminando t, presenta le stesse buone caratteristiche del Modello 1, ma in questo caso tutti i predittori risultano significativi. Inoltre R^2_{adj} è leggermente aumentato e i p-value del test di significatività della regressione è leggermente diminuito.

   Per questi motivi è opportuno scegliere il Modello 2.

3. $\widehat{H}|_{p=1.3,d=0.4} = 0.53096 + 0.4785 \times 1.3 - 0.44194 \times 0.4 = 0.976234.$

4. $\widehat{H}|_{p,d+0.1} - \widehat{H}|_{p,d} = \widehat{\beta}_2 \times 0.1 = -0.44194 \times 0.1 = -0.044194$

   quindi stimiamo che, se il diametro dello spot d aumenta di 1 e gli altri predittori non variano, il rapporto di forma H in media diminuisce di 0.044194.

5. $\mathbb{E}[H|d = 0, p = 0] = \beta_0$ per cui dobbiamo calcolare una stima intervallare al 90% per β_0, ovvero

$$\hat{\beta}_0 \pm t_{0.05,24-3}\text{se}(\hat{\beta}_0) = 0.53096 \pm 1.721 \times 0.07149 = 0.53096 \pm 0.1230343 = [0.4079257; 0.6539943].$$

6. Se $e_{(1)} = -0.103125$ e se q_α denota il quantile di ordine α di una normale standard, allora, avendo $n = 24$ prove, il corrispondente punto nel Normal Probability Plot dei residui standardizzati ha coordinate

$$\left(q_{\frac{1-0.5}{n}}, r_{(1)}\right) = \left(q_{\frac{1}{2n}}, \frac{e_{(1)}}{\widehat{\sigma}}\right) = \left(q_{0.02083}, \frac{-0.103125}{0.0592}\right) = \left(-2.037, -1.742\right)$$