Inferenza Statistica

Note in R

V. Gioia (e R. Pappadà, e N. Torelli)

18/12/2024

Contents

Esercizio 2 - Esercitazione 10

Un'indagine campionaria su 900 utenti di un provider internet ha permesso di rilevare che il 73% di essi è completamente soddisfatto del servizio. Qualche hanno dopo si riscontra che su un campione di 750 utenti il 70% è completamente soddisfatto del servizio. Si adotti una opportuna procedura di verifica di ipotesi, a livello $\alpha = 0.05$, per determinare se non vi siano variazioni nel tempo della proporzione di utenti completamente soddisfatti del servizio.

Quindi denotiamo con $x_1, \dots, x_n$, le rilevazioni sulla soddisfazione al tempo t_1 su un campione di $n = 900$ utenti di un provider internet, e con $y_1, \dots, y_m$, le rilevazioni sulla soddisfazione al tempo t_2 su un campione di $m = 750$ il campione relativo.

Essi sono realizzazioni di $X_1, \dots, X_n \overset{iid}{\sim} \text{Be}(p_X)$ e $Y_1, \dots, Y_m \overset{iid}{\sim} \text{Be}(p_Y)$ rispettivamente (ad es. $X_i = 1$ se completamente soddisfatto e $X_i = 0$ altrimenti).

Si vuole testare se non vi siano variazioni nel tempo della proporzione di utenti completamente soddisfatti del servizio, ovvero

$$\begin{cases} H_0 : p_X = p_Y & (p_X - p_Y = 0) \\ H_1 : p_X \neq p_Y & (p_X - p_Y \neq 0) \end{cases}$$

Nota: in alcuni casi può essere più sensato sottoporre a verifica se non vi sia un aumento o diminuizione nel tempo, pertanto può ad es. apparire sensato testare $H_0 : p_Y - p_X \leq 0$ contro $H_1 : p_Y - p_X > 0$ (o $H_0 : p_Y - p_X \geq 0$ contro $H_1 : p_Y - p_X < 0$).

Pertanto siano $P_X = (\sum_{i=1}^{n} X_i)/n$ e $P_Y = (\sum_{i=1}^{m} Y_i)/m$ (gli stimatori) delle proporzioni campionarie. Sappiamo che per n e m grande

$$P_X \dot{\sim} \mathcal{N}(p_X, p_X(1-p_X)/n) \qquad P_Y \dot{\sim} \mathcal{N}(p_Y, p_Y(1-p_Y)/m)$$

Quindi

$$P_X - P_Y \dot{\sim} \mathcal{N}(p_X - p_Y, p_X(1-p_X)/n + p_Y(1-p_Y)/m)$$

Sotto $H_0 : p_X = p_Y = p$,

$$P_X - P_Y \dot{\sim} \mathcal{N}(0, p(1-p)/n + p(1-p)/m)$$

e quindi necessitiamo di uno stimatore della proporzione comune

$$P_c = \frac{n}{n+m}P_X + \frac{m}{n+m}P_Y = \frac{nP_X + mP_Y}{n+m}$$

Una sua stima è data $\hat{p}_c = \frac{n\hat{p}_X + m\hat{p}_Y}{n+m} = 0.716$, avendo denotato con $\hat{p}_X = (\sum_{i=1}^{n} x_i)/n = 0.73$ e $\hat{p}_Y = (\sum_{i=1}^{n} y_i)/m = 0.7$

Quindi, sotto $H_0 : p_1 - p_2 = 0$ abbiamo

$$P_X - P_Y \dot{\sim} \mathcal{N}(0, \hat{p}_c(1-\hat{p}_c)(1/n + 1/m))$$

Per ottenere la regione di rifiuto procediamo come

$$\alpha = P(P_X - P_Y \in \mathcal{R}|H_0) = P\left(\frac{|P_X - P_Y - (p_X - p_Y)|}{\sqrt{\hat{p}_c(1-\hat{p}_c)(1/n+1/m)}} > k'|p_X - p_Y = 0\right)$$

$$= P\left(\frac{|P_X - P_Y|}{\sqrt{\hat{p}_c(1-\hat{p}_c)(1/n+1/m)}} > z_{1-\alpha/2}\right) = P\left(|P_X - P_Y| > z_{1-\alpha/2}\sqrt{\hat{p}_c(1-\hat{p}_c)(1/n+1/m)}\right)$$

$$\mathcal{R} = \{\mathbf{x}, \mathbf{y} : |\mathbf{\hat{p}_X} - \mathbf{\hat{p}_Y}| > \mathbf{z_{1-\alpha/2}}\sqrt{\mathbf{\hat{p}_c(1-\hat{p}_c)(1/n+1/m)}}\}$$

Nel nostro caso si ha con $\alpha = 0.05$, $\hat{p}_X = 0.73$ e $\hat{p}_Y = 0.70$ e $\hat{p}_c = 0.716$

$$\mathcal{R} = \{\mathbf{x_1}, \mathbf{x_2} : |\mathbf{\hat{p}_X} - \mathbf{\hat{p}_Y}| > \mathbf{1.96}\sqrt{\mathbf{0.716(1-0.716)(1/900+1/750)}} = \mathbf{0.044}\}$$

poichè $\hat{p}_X - \hat{p}_Y = 0.03 \notin \mathcal{R}$ per cui non vi è evidenza empirica per confutare che le differenze tra le proporzioni siano cambiate nel tempo.

Il p-value risulta essere

$$p - value = P\left(\frac{|P_X - P_Y - (p_X - p_Y)|}{\sqrt{\hat{p}_c(1-\hat{p}_c)(1/n+1/m)}} > \frac{|\hat{p}_X - \hat{p}_Y - (p_X - p_Y)|}{\sqrt{\hat{p}_c(1-\hat{p}_c)(1/n+1/m)}}\middle| H_0\right) =$$

$$= P\left(\frac{|P_X - P_Y|}{\sqrt{\hat{p}_c(1-\hat{p}_c)(1/n+1/m)}} > \frac{0.3}{\sqrt{\hat{p}_c(1-\hat{p}_c)(1/n+1/m)}}\right) = 2\times\left(1 - \Phi\left(\frac{0.3}{\sqrt{\hat{p}_c(1-\hat{p}_c)(1/n+1/m)}}\right)\right) =$$

$$= 2 \times (1 - \Phi(1.345598)) = 0.1784322$$

Vediamo in R cosa cambia, l'utilizzo della precedente procedura può essere fatto a mano:

- Tuttavia esiste una funzione **prop.test** la quale richiede in argomento il vettore contenente il totale di successi e il vettore contenente le numerosità campionarie dei due campioni
- Di default, viene considerato il sistema di ipotesi avente regione di rifiuto bilaterale (si può ovviamnte convertire in unilaterale, si esplori l'**help** a tal proposito)
- Dato che questo test si può adottare per piccoli campioni, esiste una correzione, pertanto impostiamo il paramtetro **correct = FALSE**, tuttavia per grandi campioni il ruolo della correzione è pressochè insignificante
- Tuttavia dall'output potreste rimanere straniti dal fatto che si riporti una statistica test **chi-squared** ma essendo alla fine del corso di inferenza, capiamo banalmente il motivo data la relazione tra una normale e una chi-quadrato

- Come potete notare si perviene alle stesse conclusioni del test riportato sopra. Ovviamente non viene fornita la regione di rifiuto, la quale richiede la specifica di un certo α mentre viene riportato soltanto il p-value
- Viene fornito l'intervallo di confidenza al 95% (di default), e si ricordi la dualità tra intervalli di confidenza e test d'ipotesi

```
n <- 900
m <- 750

x <- c(rep(0, 0.27 * 900), rep(1, 0.73 * 900))
y <- c(rep(0, 0.3 * 750), rep(1, 0.70 * 750))

data <- c(sum(x), sum(y))

prop.test(data, c(n,m), correct = FALSE)
```

```
##
##  2-sample test for equality of proportions without continuity correction
##
## data:  data out of c(n, m)
## X-squared = 1.812, df = 1, p-value = 0.1783
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.01378228  0.07378228
## sample estimates:
## prop 1 prop 2
##   0.73   0.70
```

```
pc <- (mean(x) * n + mean(y) * m)/(n + m)

test_stat <- (mean(x) - mean(y))/(sqrt(pc * (1 -pc) * (1/n + 1/m)))
test_stat^2
```

```
## [1] 1.812036
```

```
2 * pnorm(abs(test_stat), lower = FALSE)
```

```
## [1] 0.1782641
```

```
pchisq(test_stat^2, df = 1, lower = FALSE)
```

```
## [1] 0.1782641
```

Test di Kolmogorov-Smirnov

Dato un campione casuale $y_1, \dots, y_n$ proveniente da Y, con f.d.r. $F(y)$. Si denota con $F_n(y)$ la f.d.r. empirica

$$F_n(y) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{I}_{\{-\infty, y\}}(y_i)$$

Si vuole sottoporre a test

$$\begin{cases} H_0 : F(y) = F_0(y) \\ H_1 : F(y) \neq F_0(y) \end{cases}$$

dove $F_0(y)$ definisce una legge continua su $\mathbb{R}$. Il test di Kolmogorov-Smirnov si basa su

$$D_n = \sup_{\{-\infty < y < \infty\}} |F_n(y) - F_0(y)|$$

Pertanto la regione critica risulterà

$$\mathcal{R} = \{(y_1, \dots, y_n) : \sup_{\{-\infty < y < \infty\}} |F_n(y) - F_0(y)| > d_{1-\alpha}\}$$

con d_{1-a} soglia critica per il test ed è tabulata per dimensioni non elevate. Operativamente

- Si ordinano le osservazioni $y_1, \dots, y_n$ in senso crescente, ottenendo $y_{(1)}, \dots, y_{(n)}$
- Si ottengono $F_0(y_{(i)})$, $i = 1, \dots, n$, e $F_n(y_{(i)}) = i/n$ (in presenza di dati assolutamente continui; in presenza di due valori identici, i cosiddetti ties, bisogna considerare la definizione)
- Si calcolano gli scostamenti della funzione di ripartizione empirica dalla funzione di ripartizione ipotizzata

$$D_n^+ = \max_{\{1 \leq i \leq n\}} \left(F_n(y_{(i)}) - F_0(y_{(i)}) \right) = \max_{\{1 \leq i \leq n\}} \left(i/n - F_0(y_{(i)}) \right)$$

$$D_n^- = \max_{\{1 \leq i \leq n\}} \left(F_0(y_{(i)}) - F_n(y_{(i-1)}) \right) = \max_{\{1 \leq i \leq n\}} \left(F_0(y_{(i)}) - (i-1)/n \right)$$

$$D_n = \max(D_n^+, D_n^-)$$

Esempio Il peso espresso in Kg di 6 corridori/corritrice è dato da $\mathbf{y} = (58.1, 55.2, 59.7, 62.8, 64.2, 54.5)$. Verificare a un livello di significatività pari a 0.1 se il peso Y si possa ritenere distribuito con legge normale di media 60kg e deviazione standard di 4kg utilizzando il test di Kolmogorov-Smirnov.

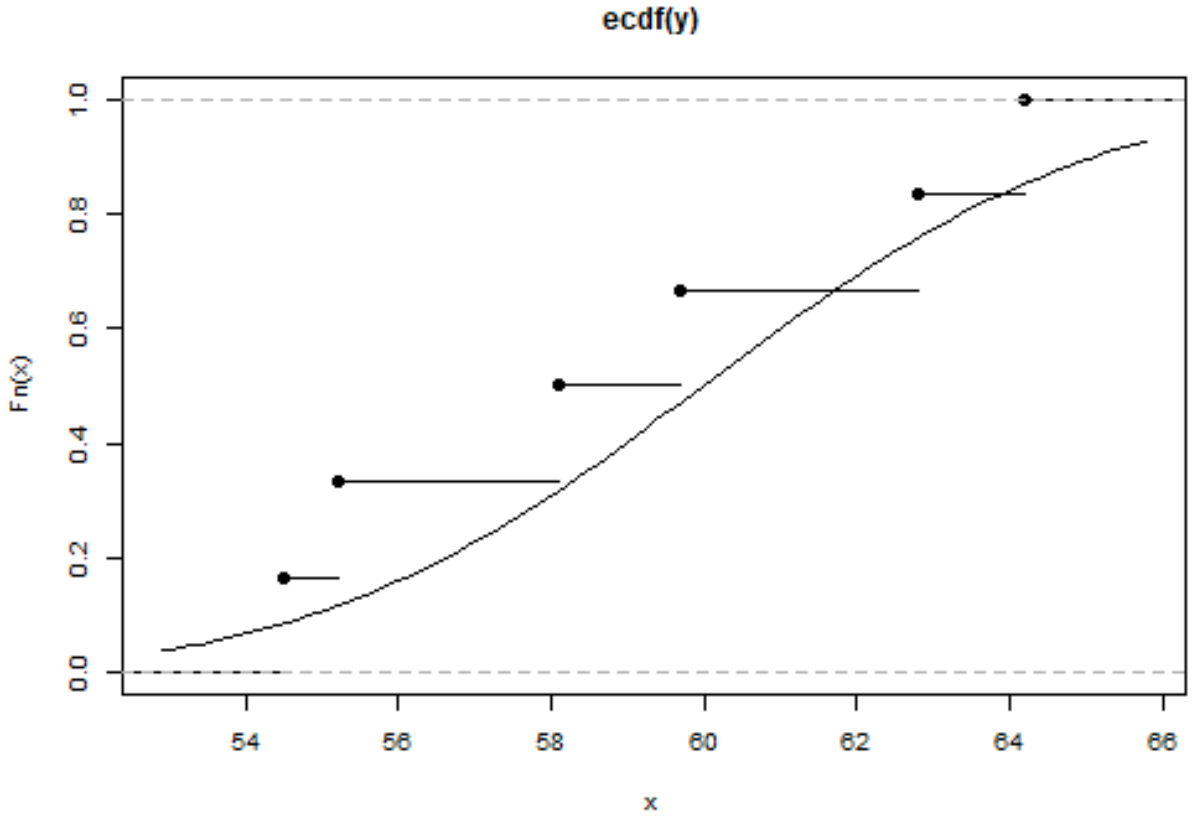
Procediamo prima per via analitica.

i	$y_{(i)}$	$F_n(y_{(i)})$	$F_0(y_{(i)})$	$D_n^+ = F_n(y_{(i)}) - F_0(y_{(i)})$	$D_n^- = F_0(y_{(i)}) - F_n(y_{(i-1)})$
1	54.5	1/6	0.0846	0.0820	-0.0820
2	55.2	2/6	0.1151	0.2183	-0.0516
3	58.1	3/6	0.3174	0.1826	-0.0159
4	59.7	4/6	0.4701	0.1966	-0.0299
5	62.8	5/6	0.7580	0.0753	0.0914
6	64.2	1	0.8531	0.1469	0.0198

da cui $D_n = max(D_n^+, D_n^-) = 0.2183$. In R.

```
y <- c(58.1, 55.2, 59.7, 62.8, 64.2, 54.5)

plot(ecdf(y))
curve(pnorm(x, 60, 4), add =TRUE)
```




```
ks.test(y, "pnorm", 60, 4)
```

```
##
##  Exact one-sample Kolmogorov-Smirnov test
##
## data:  y
## D = 0.21826, p-value = 0.8837
## alternative hypothesis: two-sided
```

Il calcolo a mano della statistica test si può ottenere come

```
ys <- sort(y)
Fn <- c(0, (1:length(ys))/length(ys))
F0 <- pnorm(ys, 60, 4)
max(Fn[-1] - F0, F0 - Fn[-length(Fn)])
```

```
## [1] 0.2182637
```

mentre il calcolo del valore p è più complesso (si veda l'*help* della funzione *ks.test*()). Tuttavia volendo prendere una decisione si può confrontare tale valore con quello tabulato per cui risulta $d_{1-\alpha;6} = d_{0.9;6} = 0.468$. Dal momento che $D_n \notin \mathcal{R}$ non vi è evidenza empirica per confutare H_0.

Come cambiano le conclusioni se

- Se avessimo assunto che il campione provenisse da una $\mathcal{N}(60, \sigma^2 = 2)$, **qui le conclusioni non cambiano**

```
ks.test(y, "pnorm", 60, 2)
```

```
##
##  Exact one-sample Kolmogorov-Smirnov test
##
## data:  y
## D = 0.32894, p-value = 0.4392
## alternative hypothesis: two-sided
```

- Se avessimo assunto che il campione provenisse da una $\mathcal{N}(65, \sigma^2 = 16)$, **qui le conclusioni cambiano in favore di H_1**

```
ks.test(y, "pnorm", 65, 4)
```

```
##
##  Exact one-sample Kolmogorov-Smirnov test
##
## data:  y
## D = 0.57926, p-value = 0.01938
## alternative hypothesis: two-sided
```

Esercizio Si provi a generare un campione di dimensione $n = 50$ da una distribuzione *Gamma* di parametri $\lambda = 2$ e $\alpha = 3$ e si sottoponga a test l'assunzione che i dati vengano da tale distribuzione.

Test di indipendenza

Consideriamo il seguente esempio per verificare se due caratteri di natura qualitativa sono indipendenti o meno. La classificazione del gruppo sanguigno ABO prevede quattro possibili gruppi A, B, AB e O. Al fine di verificare se la distribuzione del gruppo sanguigno sia indipendente dalla macro regione italiana di appartenenza ("Sud e Isole", "Centro", "Nord"), si considera un campione di 760 individui. La seguente tabella mostra la distribuzione congiunta del gruppo sanguigno X e della macro-regione di appartenenza Y

	A	B	AB	O
SUD e ISOLE	50	70	30	100
CENTRO	114	30	10	100
NORD	116	27	13	100

Quindi osserviamo la tabella delle frequenze assolute

	x_1	x_2	x_3	x_4	Total
y_1	n_{11}	n_{12}	n_{13}	n_{14}	$n_{1\cdot}$
y_2	n_{21}	n_{22}	n_{23}	n_{24}	$n_{2\cdot}$
y_3	n_{31}	n_{32}	n_{33}	n_{34}	$n_{3\cdot}$
Total	$n_{\cdot 1}$	$n_{\cdot 2}$	$n_{\cdot 3}$	$n_{\cdot 4}$	n

da cui si deriva quelle delle frequenze relative $f_{ij} = n_{ij}/n$, $f_{i\cdot} = n_{i\cdot}/n$ e $f_{\cdot j} = n_{\cdot j}/n$.

	x_1	x_2	x_3	x_4	Total
y_1	f_{11}	f_{12}	f_{13}	f_{14}	$f_{1\cdot}$
y_2	f_{21}	f_{22}	f_{23}	f_{24}	$f_{2\cdot}$
y_3	f_{31}	f_{32}	f_{33}	f_{34}	$f_{3\cdot}$
Total	$f_{\cdot 1}$	$f_{\cdot 2}$	$f_{\cdot 3}$	$f_{\cdot 4}$	1

Quindi si vuole sottoporre a verifica il seguente sistema di ipotesi

$$\begin{cases} H_0 : f_{ij} = f_{i\cdot} f_{\cdot j} \ \ \forall i = 1, \dots, s, \ j = 1, \dots, t \\ H_1 : \exists \text{ almeno } (i, j) : f_{ij} \neq f_{i\cdot} f_{\cdot j} \end{cases}$$

```
n <- 760
obs_freq <- matrix(c(50, 70, 30, 100,
                     114, 30, 10, 100,
                     116, 27, 13, 100),3,4,T)
colnames(obs_freq) <- c("A", "B", "AB", "O")
rownames(obs_freq) <- c("Sud", "Centro", "Nord")
chisq.test(obs_freq)
```

```
##
##  Pearson's Chi-squared test
##
## data:  obs_freq
## X-squared = 70.99, df = 6, p-value = 2.561e-13
```

Di seguito i passaggi manuali per ottenere i risultati della procedura **chisq.test**.

```
mx <- colSums(obs_freq)/n; mx
```

```
##          A          B         AB          O
## 0.36842105 0.16710526 0.06973684 0.39473684
```

```
my <- rowSums(obs_freq)/n; my
```

```
##       Sud    Centro      Nord
## 0.3289474 0.3342105 0.3368421
```

```r
exp_freq <- outer(my, mx) * n; exp_freq
```

```
##                A        B       AB         O
## Sud    92.10526 41.77632 17.43421  98.68421
## Centro 93.57895 42.44474 17.71316 100.26316
## Nord   94.31579 42.77895 17.85263 101.05263
```

```r
chi2 <- sum((obs_freq - exp_freq)^2/exp_freq); chi2
```

```
## [1] 70.99012
```

```r
pchisq(chi2, (ncol(obs_freq) - 1) * (nrow(obs_freq) - 1), lower.tail = FALSE)
```

```
## [1] 2.561274e-13
```

Si noti che rimuovendo la macro-regione "Sud e Isole" l'indicazione del test è sostanzialmente diversa. **Esercizio**: Si conduca l'analogo test non considerando la ripartizione "Sud e Isole".

Test di conformità

Supponiamo di voler verificare l'ipotesi che un dado non sia truccato

$$\begin{cases} H_0 : \pi_1 = \pi_2 = \pi_3 = \pi_4 = \pi_5 = \pi_6 = 1/6 \\ H_1 : \exists\, k = 1, \dots, 6 : \pi_k \neq 1/6 \end{cases}$$

Denotiamo con

- $n_j = \sum_{i=1}^{n} \mathbf{1}_{\{y_i = j\}}$, le frequenze assolute osservate
- $\hat{n}_j = n\pi_j = n/6$ le frequenze assolute attese sotto H_0

Quindi calcoliamo

$$X^2 = \sum_{j=1}^{6} \frac{(n_j - \hat{n}_j)^2}{\hat{n}_j}$$

Se fosse vera H_0 (il dado non è truccato) tale statistica dovrebbe essere vicina a 0: ci si attende che l'1 si presenti circa 1/6 di volte, lo stesso varrà per il 2, etc.. Le differenze tra frequenze osservate e frequenze teoriche saranno piccole, dovute al più alle fluttuazioni casuali. I valori grandi di X^2 saranno allora quelli sospetti e dovremo fissare la soglia oltre la quale decideremo che il valore ottenuto è sufficientemente grande da potere ritenere ragionevole rifiutare H_0. Nel nostro caso, essendo $K = 6$ le modalità, X^2 ha distribuzione limite, sotto H_0, χ^2 con $k - 1 = 5$ gradi di libertà.

```r
set.seed(10)
n <- 200
dado <- sample(1:6, n, replace = TRUE)
freq <- table(dado); freq
```

```
## dado
##  1  2  3  4  5  6
## 31 34 31 28 36 40
```

```r
chisq.test(freq)
```

```
##
##  Chi-squared test for given probabilities
##
## data:  freq
## X-squared = 2.74, df = 5, p-value = 0.74
```

```r
# A mano
attese <- n * (1/6)
chi_sq <- sum((freq - attese)^2/attese)
chi_sq
```

```
## [1] 2.74
```

```r
qchisq(0.95, 5)
```

```
## [1] 11.0705
```

```r
pchisq(chi_sq, 5, lower.tail = FALSE)
```

```
## [1] 0.7399941
```

Consideriamo ora una generazione casuale che provi ad emulare il lancio ripetuto di un dado truccato. Conducendo il test vi è evidenza empirica per confutare H_0.

```r
dado_truc <- sample(1 : 6, n, p=c(0.1,0.1, 0.30,0.30,0.1,0.1), replace = TRUE)
freq <- table(dado_truc); freq
```

```
## dado_truc
##  1  2  3  4  5  6
## 25 21 60 55 20 19
```

```r
chisq.test(freq)
```

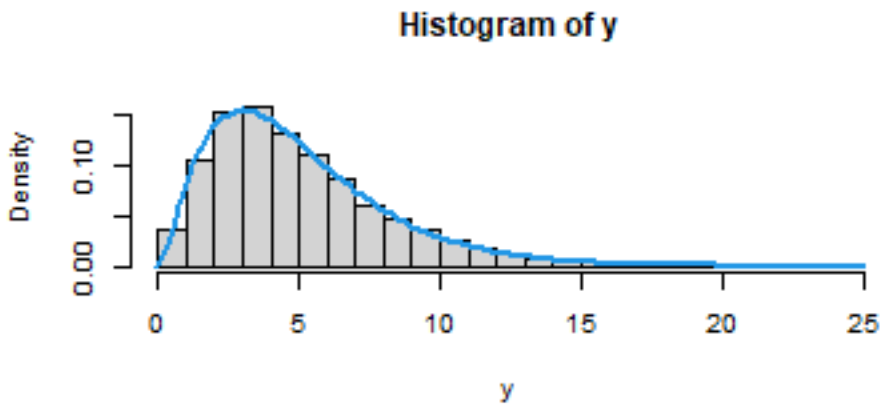
```
##
##  Chi-squared test for given probabilities
##
## data:  freq
## X-squared = 53.56, df = 5, p-value = 2.581e-10
```

Conduciamo quindi una simulazione Monte Carlo. Per ogni replicazione ($R = 1000$)

- Simuliamo il lancio di un dado un certo numero di volte ($n = 100$)
- Calcoliamo la statistica test del test X^2 di conformità

In questo modo siamo in grado di ottenere la distribuzione della statistica test e possiamo confrontarla con quella teorica (χ^2_5)

```r
R <- 10000
attese <- n * (1/6)
y <- vector(mode = "numeric", length = R)
for (i in 1 : R) {
  dado <- sample(1 : 6, n, replace = TRUE)
  freq <- table(dado)
  y[i] <- sum((freq-attese)^2/attese)
}
hist(y, prob = T, breaks = 30, ylim = c(0,0.17))
curve(dchisq(x, df = 5), type = "l", col = 4, add = T, lwd = 2)
```



Si può vedere quindi che ripetendo l'esperimento un numero elevato di volte si ottiene che il tasso di rifiuto empirico è vicino al livello nominale α, ovvero la probabilità di errore del I tipo.

```r
sum(y >= qchisq(0.95, 5))/R
```

```
## [1] 0.0496
```

Esercizio 2 - Esercitazione 11

Si osserva un campione casuale di 3 valori, x_1, x_2, x_3, da $X \sim \text{Poisson}(\lambda)$. Si ottenga la regione di rifiuto per il test più potente per verificare l'ipotesi $H_0 : \lambda = \lambda_0 = 1$ contro l'alternativa $H_1 : \lambda = \lambda_1 = 2$ a livello del 0.4% (si approssimino i valori della funzione di ripartizione alla terza decimale).

Durante l'esercitazione si era trovato che il test più potente aveva come regione di rifiuto a livello $\alpha = 0.004$

$$R = \{(x_1, x_2, x_3) : \sum_{i=1}^{3} x_i > 8\}$$

A tal proposito mostriamo mediante simulazione montecarlo che il tasso di rifiuto è vicino al valore nominale $\alpha = 0.004$. Si ricorda che il quantile di ordine 0.996 di una Poisson era 8.

```
ppois(8,3)
```

```
## [1] 0.996197
```

```
ppois(9,3)
```

```
## [1] 0.9988975
```

```
set.seed(1)
R <- 1000000
n <- 3
stat <- rep(NA, R)
for(i in 1:R){
 sample <- rpois(3, 1)
 stat[i] <- sum(sample)
}

sum(stat > 8)/R
```

```
## [1] 0.003838
```

```
1 - ppois(8,3)
```

```
## [1] 0.003802992
```