Verifica del modello: test di adattamento

8 giugno 2017

Verifica del modello

Problema: riconoscere quando un modello probabilistico si adatta ad un fenomeno casuale.

Esempio 1. In una roulette ci sono 18 numeri rossi, 18 numeri neri e lo zero. La roulette è non truccata se tutti i 37 numeri hanno la stessa probabilità di uscire in una giocata.

Sto giocando ad una roulette.
Su $n = 300$ giocate:
135 volte è uscito un numero rosso;
159 volte è uscito un numero nero;
6 volte è uscito lo zero.

Posso concludere che la roulette è truccata?

Esempio 2. In una ricerca condotta qualche anno fa, il numero di incidenti automobilistici per settimana in un tratto di autostrada seguiva una legge di Poisson di parametro $\lambda = 0.4$. Se nelle ultime 85 settimane si sono rilevati i seguenti dati

N^o incidenti sett.:	0	1	2	≥ 3	$\sum$
N^o settimane in cui si è verificato:	50	32	3	0	85

Si può affermare che il modello ipotizzato ($Poiss(0.4)$) non è più applicabile alla descrizione del fenomeno?

Esempio 3. Con un generatore di numeri casuali che simula una popolazione U uniformemente distribuita nell'intervallo [0,1] (cioè U v.a. tale che $U \sim \mathcal{U}(0,1)$), sono stati generati 150 numeri $u_1, \ldots, u_{150}$ ottenendo:

intervallo:	$[0, 1/4]$	$(1/4, 1/2]$	$(1/2, 3/4]$	$[3/4, 1]$
N^o dati nell'intervallo:	30	37	52	31

Verificare l'ipotesi nulla "il generatore è veramente casuale", cioè che $u_1, \ldots, u_{150}$ siano valori assunti da $U_1, \ldots, U_{150}$ campione aleatorio estratto da una popolazione uniforme su $[0, 1]$.

Test di buon adattamento

I test statistici che servono a verificare se un modello probabilistico è compatibile con i dati sono detti test di buon adattamento.

(goodness of fit tests)

Test χ^2 di buon adattamento

Un'ampia popolazione è formata da oggetti che possono essere classificati in k categorie o classi diverse.

Per $i = 1, \ldots, k$, sia p_i la probabilità che un oggetto scelto a caso dalla popolazione appartenga alla classe i. Quindi

$$p_i \geq 0 \quad \text{e} \quad \sum_{i=1}^{k} p_i = 1$$

Siano $p_1^0, \ldots, p_k^0$ k numeri assegnati tali che

$$p_i^0 > 0 \quad \text{e} \quad \sum_{i=1}^{k} p_i^0 = 1.$$

Esempio 1.

- Popolazione: numeri che escono alla roulette.
- Categorie k=3: 1=ROSSO, 2=NERO, 3=ZERO.
  $p_1 = P(\text{ROSSO})$, $p_2 = P(\text{NERO})$, $p_3 = P(\text{ZERO})$.
- Modello roulette non truccata: $p_1^0 = 18/37$, $p_2^0 = 18/37$, $p_3^0 = 1/37$.

Vogliamo verificare:

$\mathbb{H}_0 : p_i = p_i^0$ per $i = 1, \ldots, k$
contro l'ipotesi alternativa
$\mathbb{H}_1 : p_i \neq p_i^0$ per almeno un i.

Consideriamo un campione di dimensione n estratto dalla popolazione, cioè n osservazioni indipendenti e ciascuna ha probabilità p_i di appartenere alla categoria/classe i.
Sia, per $i = 1, \ldots, k$,
X_i= n^0 delle osservazioni nel campione che sono nella classe i.
(frequenza assoluta nel campione della classe i)

$\Rightarrow X_i \geq 0$ per $i = 1, \ldots, k$, $\sum_{i=1}^{k} X_i = n$
$X_i \sim Bin(n, p_i)$ e quindi $\mathbb{E}(X_i) = np_i$ per $i = 1, \ldots k$.

N.B. X_i può essere visto come n^0 dei successi (appartenenza alla classe i) in n prove di Bernoulli (osservazioni nel campione).

Idea del test

Poiché $X_i \sim Bin(n, p_i)$ e $\mathbb{E}(X_i) = np_i$ per $i = 1, \dots k$, segue che

$$(X_i - np_i^0)^2$$

è un indicatore di quanto sia verosimile $p_i = p_i^0$: quando queste quantità sono troppo grandi mi suggeriscono un rifiuto di $\mathbb{H}_0$.
Quindi è naturale usare come statistica-test:

$$T := \sum_{i=1}^{k} \frac{(X_i - np_i^0)^2}{np_i^0}.$$

L'ipotesi nulla $\mathbb{H}_0$ va rifiutata quando $T \geq c$, c determinata in base al livello di significatività α del test:

$$P_{\mathbb{H}_0}(T \geq c) = \alpha.$$

Se n "grande" e sotto $\mathbb{H}_0$ $T := \sum_{i=1}^{k} \frac{(X_i - np_i^0)^2}{np_i^0} \approx \chi^2(k-1) \Rightarrow c \simeq \chi^2_{\alpha,k-1}$.

In conclusione: Sia

$$T := \sum_{i=1}^{k} \frac{(X_i - np_i^0)^2}{np_i^0},$$

dove, per $i = 1, \ldots, k$, $X_i = n^0$ delle osservazioni nel campione che appartengono alla classe i.
Per verificare:

$\mathbb{H}_0 : p_i = p_i^0$ per $i = 1, \ldots, k$
contro l'ipotesi alternativa
$\mathbb{H}_1 : p_i \neq p_i^0$ per almeno un i.

un test di livello approssimato α è quello che, osservato $T = \bar{t}$

rifiuta $\mathbb{H}_0$ se $\bar{t} \geq \chi^2_{\alpha,k-1}$

accetta $\mathbb{H}_0$ se $\bar{t} < \chi^2_{\alpha,k-1}$.

$$p-value = P_{\mathbb{H}_0}(T \geq \bar{t}) \simeq P(W \geq \bar{t})$$

dove $W \sim \chi^2(k-1)$ (k = numero delle classi/categorie).

Osservazioni

1. Regola empirica per stabilire quando n è "abbastanza grande": almeno l'80% delle $np_i^0 \geq 5$ e le rimanenti $np_i^0 > 1$.
2. Formula utile: ricorda $\sum_{i=1}^k p_i^0 = 1$ e $\sum_{i=1}^k X_i = n$

$$\Downarrow$$

$$T = \sum_{i=1}^k \frac{(X_i - np_i^0)^2}{np_i^0} = \sum_{i=1}^k \frac{X_i^2 + n^2(p_i^0)^2 - 2np_i^0 X_i}{np_i^0}$$

$$= \sum_{i=1}^k \frac{X_i^2}{np_i^0} + n\sum_{i=1}^k p_i^0 - 2\sum_{i=1}^k X_i$$

$$= \sum_{i=1}^k \frac{X_i^2}{np_i^0} - n.$$

3 Sia $k = 2 \Rightarrow X_1 + X_2 = n$ e $p_1^0 + p_2^0 = 1$.

$$
\begin{aligned}
T &= \frac{(X_1 - np_1^0)^2}{np_1^0} + \frac{(X_2 - np_2^0)^2}{np_2^0} \\
&= \frac{(X_1 - np_1^0)^2}{np_1^0} + \frac{\left(n - X_1 - n(1 - p_1^0)\right)^2}{n(1 - p_1^0)} \\
&= \frac{(X_1 - np_1^0)^2}{np_1^0} + \frac{(X_1 - np_1^0)^2}{n(1 - p_1^0)} = \frac{(X_1 - np_1^0)^2}{np_1^0(1 - p_1^0)}.
\end{aligned}
$$

Se "n grande e $\mathbb{H}_0$ è vera " $\Rightarrow$ $\dfrac{X_1 - np_1^0}{\sqrt{np_1^0(1 - p_1^0)}} \approx \mathcal{N}(0, 1)$

e quindi $T = \dfrac{(X_1 - np_1^0)^2}{np_1^0(1 - p_1^0)} \approx \chi^2(1)$

4. Se $k > 2$, la dimostrazione che la distribuzione asintotica della statistica test è chi-quadrato con $k - 1$ gradi di libertà, in simboli:

$$\text{per } n \text{ “grande”, sotto } \mathbb{H}_0 \Rightarrow T = \sum_{i=1}^{k} \frac{(X_i - np_i^0)^2}{np_i^0} \approx \chi^2(k-1)$$

è più difficile. La statistica test è funzione delle k v.a. $X_1, \ldots, X_k$ tali che $\sum_{i=1}^{k} X_i = n$. Il legame tra le X_i fa “perdere un grado di libertà”: sotto $\mathbb{H}_0$ $T \approx \chi^2(k-1)$.

Esempio 1 (roulette). Dati: $n = 300$ giocate, 135 volte esce un numero rosso, 159 volte esce uno nero e 6 volte esce lo zero.

Sia $p_1 = P(\text{rosso})$, $p_2 = P(\text{nero})$ e $p_3 = P(\text{zero})$

$\mathbb{H}_0$: la roulette non è truccata: $p_1 = \frac{18}{37}$, $p_2 = \frac{18}{37}$, $p_3 = \frac{1}{37}$.

contro l'ipotesi alternativa

$\mathbb{H}_1$: la roulette è truccata.

($p_1^0 = \frac{18}{37}$, $p_2^0 = \frac{18}{37}$, e $p_3^0 = \frac{1}{37}$.)

classe i:	$1 =$ rosso	$2 =$ nero	$3 =$ zero	$\sum = n$
X_i	135	159	6	300
np_i^0	145.946	145.946	8.108	300

La statistica-test $T := \sum_{i=1}^3 \frac{(X_i - np_i^0)^2}{np_i^0} = \sum_{i=1}^3 \frac{X_i^2}{np_i^0} - n \simeq 2.54$

$p - value \simeq P(W \geq 2.54) \simeq 0.2813$ dove $W \sim \chi^2(k-1) = \chi^2(2)$.

Non posso rifiutare $\mathbb{H}_0$ agli usuali livelli di significatività con i dati a disposizione.

N.B. $\chi^2_{0.1,2} = 4.61$, $\chi^2_{0.05,2} = 5.991$, $\chi^2_{0.01,2} = 9.210$.

Esempio 2. I dati:

classe i	classe 1$\rightarrow$ $\{0$ incid.$\}$	classe 2$\rightarrow$ $\{1$ incid.$\}$	classe 3$\rightarrow$ $\{2$ incid.$\}$	classe 4$\rightarrow$ $\{\geq 3$ incid.$\}$	$\sum = n$
N^o sett. in cui si è verificato	50	32	3	0	85

Il problema di verifica d'ipotesi può essere cosí formalizzato:

$Y =$ N^o incidenti in una settimana nel tratto autostradale

$F :=$ distribuzione di Y

Vogliamo verificare, sulla base di questi dati

$\mathbb{H}_0 : F = Poiss(0.4)$ contro $\mathbb{H}_1 : F \neq Poiss(0.4)$.

Sia $Y_1, \ldots, Y_{85}$ campione aleatorio di dimensione $n = 85$, dove $Y_j = N^o$ incidenti nella settimana j nel tratto autostradale.

Non conosciamo una realizzazione campionaria $(y_1, \ldots, y_{85})$

Conosciamo (x_1, x_2, x_3, x_4) dove, per $i = 1, 2, 3, 4$, x_i è il valore di $X_i = N^o$ degli elementi del campione che appartengono alla classe i

$=$ frequenza della classe i.

Riassumendo i nostri dati sono i valori assunti dalle frequenze X_1, X_2, X_3, X_4, nel campione delle classi:

$\{0 \text{ incidenti}\}$, $\{1 \text{ incidenti}\}$, $\{2 \text{ incidenti}\}$, $\{\geq 3 \text{ incidenti}\}$.

Sia, per $i = 1, 2, 3, 4$, p_i la probabilità che un elemento del campione appartenga alla i-esima classe, cioè:

$p_1 = P(Y = 0)$, $p_2 = P(Y = 1)$, $p_3 = P(Y = 2)$, $p_4 = P(Y \geq 3)$

Queste probabilità, sotto $\mathbb{H}_0$: $Y \sim F = Poiss(0.4)$ valgono:

$$p_1^0 = P_{\mathbb{H}_0}(Y = 0) = e^{-0.4} \simeq 0.67 \quad p_2^0 = P_{\mathbb{H}_0}(Y = 1) = 0.4e^{-0.4} \simeq 0.268$$

$$p_3^0 = P_{\mathbb{H}_0}(Y = 2) = \frac{(0.4)^2 e^{-0.4}}{2} \simeq 0.054 \quad p_4^0 = 1 - p_1^0 - p_2^0 - p_3^0 \simeq 0.008$$

Quindi posso riformulare il problema di verifica d'ipotesi:

$\mathbb{H}_0$: $p_1 = 0.67$, $p_2 = 0.268$, $p_3 = 0.054$, $p_4 = 0.008$.

contro l'ipotesi alternativa

$\mathbb{H}_1$: non è vera $\mathbb{H}_0$.

ed effettuare un test χ^2 di adattamento.

classe i	classe 1$\rightarrow$ $\{0$ incid.$\}$	classe 2$\rightarrow$ $\{1$ incid.$\}$	classe 3$\rightarrow$ $\{2$ incid.$\}$	classe 4$\rightarrow$ $\{\geq 3$ incid.$\}$	$\sum = n$
X_i=N^osett. in cui si è verificato	50	32	3	0	85
np_i^0	56.95	22.78	4.59	0.68	85

Per le ultime due classi $np_i^0 < 5$ quindi le raggruppo in un'unica classe.

classe i	classe 1$\rightarrow$ $\{0$ incid.$\}$	classe 2$\rightarrow$ $\{1$ incid.$\}$	classe 3$\rightarrow$ $\{\geq 2$ incid.$\}$	$\sum = n$
X_i=N^o sett. in cui si è verificato	50	32	3	85
np_i^0	56.95	22.78	5.27	85

Il valore assunto dalla statistica-test è:

$T := \sum_{i=1}^{3} \frac{(X_i - np_i^0)^2}{np_i^0} = \sum_{i=1}^{3} \frac{X_i^2}{np_i^0} - n \simeq 5.5576$, sotto $\mathbb{H}_0$ $T \approx \chi^2(2)$ e

$p-value = P_{\mathbb{H}_0}(T \geq 5.5576) \simeq \exp(-5.5576/2) \simeq 0.0621$.

Quindi rifiuto al 10%, ma non rifiuto al 5% con questi dati.

Test di adattamento ad una distribuzione specificata a meno di parametri.

I dati:

classe i	classe 1$\rightarrow$ $\{0$ incid.$\}$	classe 2$\rightarrow$ $\{1$ incid.$\}$	classe 3$\rightarrow$ $\{2$ incid.$\}$	classe 4$\rightarrow$ $\{\geq 3$ incid.$\}$	$\sum = n$
N^o sett. in cui si è verificato	50	32	3	0	85

Consideriamo ora il problema di verifica d'ipotesi

$Y =$ N^o incidenti in una settimana nel tratto autostradale

$F :=$ distribuzione di Y

Vogliamo verificare, sulla base di questi dati

$\mathbb{H}_0 : \ F = Poisson$ contro $\mathbb{H}_1 : \ F \neq Poisson$.

L'ipotesi $\mathbb{H}_0$ non specifica il valore del parametro λ della distribuzione di Poisson.

Quindi sotto l'ipotesi $\mathbb{H}_0$ le probabilità che un campione appartenga alle classi in cui sono stati suddivisi i dati

$\{0 \text{ incidenti}\}$, $\{1 \text{ incidenti}\}$, $\{2 \text{ incidenti}\}$, $\{\geq 3 \text{ incidenti}\}$.

sono

$$p_1^0 = P_{\mathbb{H}_0}(Y = 0) = e^{-\lambda} \quad p_2^0 = P_{\mathbb{H}_0}(Y = 1) = \lambda e^{-\lambda}$$
$$p_3^0 = P_{\mathbb{H}_0}(Y = 2) = \frac{\lambda^2 e^{-\lambda}}{2} \quad p_4^0 = 1 - p_1^0 - p_2^0 - p_3^0$$

Quindi p_1^0, p_2^0, p_3^0 e p_4^0 sono specificate a meno del parametro incognito λ.

Procedimento:

- Si procede ad una stima $\hat{\lambda}$ di λ sulla base dei dati. Per esempio $\hat{\lambda}$=38/85
- Si sostituisce la stima $\hat{\lambda}$ di λ nelle espressioni delle p_i^0 ottenendo $\hat{p}_i$ per $i = 1, \ldots, 4$
- Si considera la statistica

  $\tilde{T} := \sum_{i=1}^{k} \frac{(X_i - n\hat{p}_i)^2}{n\hat{p}_i}$ dove X_i= frequenza nel campione della classe i.
- Si può dimostrare che, se n è "grande" e sotto $\mathbb{H}_0$

  $\tilde{T} := \sum_{i=1}^{k} \frac{(X_i - n\hat{p}_i)^2}{n\hat{p}_i} \approx \chi^2(k - 1 - 1)$

$k = 4$=numero delle classi, 1=numero dei parametri da stimare

un test di livello approssimato α è quello che, osservato $\tilde{T} = \tilde{t}$

rifiuta $\mathbb{H}_0$ se $\tilde{t} \geq \chi^2_{\alpha,k-2}$

accetta $\mathbb{H}_0$ se $\tilde{t} < \chi^2_{\alpha,k-2}$.

Test di adattamento per dati continui

Sia $Y_1, \ldots, Y_n$ un campione aleatorio estratto da una distribuzione F continua. Sia F_0 una distribuzione continua assegnata.
Vogliamo verificare, sulla base del campione

$$\mathbb{H}_0 : \ F = F_0 \quad \text{contro} \quad \mathbb{H}_1 : \ F \neq F_0.$$

Un approccio è: dividere i possibili valori delle v.a. Y_j, sotto $\mathbb{H}_0$, in k intervalli disgiunti

$$(y_0, y_1], \ \ldots, \ (y_{k-1}, y_k]$$

Esempi. a) Se $F_0 = \mathcal{U}(0,1) \Rightarrow 0 = y_0 < y_1 < \cdots < y_k = 1$. b) Se $F_0 = \mathcal{N}(0,1) \Rightarrow -\infty = y_0 < y_1 < \cdots < y_k = +\infty$.

I k intervalli $(y_{i-1}, y_i]$, $i = 1, \ldots, k$ sono le k classi in cui suddivideremo le osservazioni.

Quindi le k classi sono:

$$(y_0, y_1], \ \ldots, \ (y_{k-1}, y_k]$$

Le probabilità con cui le osservazioni assumono valori in queste classi:

$$p_i = P\big(Y_j \in (y_{i-1}, y_i]\big) \qquad i = 1, \ldots, k$$

(Y_j è un elemento qualsiasi del campione) e il valore di queste probabilità, sotto $\mathbb{H}_0$:

$$p_i^0 = P_{\mathbb{H}_0}\big(Y_j \in (y_{i-1}, y_i]\big) = F_0(y_i) - F_0(y_{i-1}) \qquad i = 1, \ldots, k.$$

Quindi possiamo riformulare il problema di verifica d'ipotesi:

$\mathbb{H}_0 : p_i = p_i^0$ per $i = 1, \ldots, k$
contro l'ipotesi alternativa
$\mathbb{H}_1 : p_i \neq p_i^0$ per almeno un i.

...e procedere con un test χ^2 di adattamento usuale.

Esempio 3. Generatore casuale.

$U_1 = u_1, \ldots, U_{150} = u_{150}$ $\quad n = 150$

$\mathbb{H}_0 : F = F_0 = \mathcal{U}(0,1)$

$0 = y_0 < y_1 = \frac{1}{4} < y_2 = \frac{1}{2} < y_3 = \frac{3}{4} < y_4 = 1 \Rightarrow [0, \frac{1}{4}], (\frac{1}{4}, \frac{1}{2}], (\frac{1}{2}, \frac{3}{4}], (\frac{3}{4}, 1]$

$p_i^0 = P_{\mathbb{H}_0}(y_{i-1} < U_j \leq y_i) = F_0(y_i) - F_0(y_{i-1}) = \frac{1}{4} \quad i = 1,2,3,4 \quad (k = 4).$

classe i: intervallo	$[0, 1/4]$	$(1/4, 1/2]$	$(1/2, 3/4]$	$(3/4, 1]$
$X_i = N^o$ dati nella classe i:	30	37	52	31
np_i^0:	$\frac{150}{4}$	$\frac{150}{4}$	$\frac{150}{4}$	$\frac{150}{4}$

Statistica-test: $T := \sum_{i=1}^{4} \frac{(X_i - np_i^0)^2}{np_i^0} = \frac{206}{25} = 8.24$

$p - value = P_{\mathbb{H}_0}(T \geq 8.24) \approx P(W \geq 8.24) \simeq 0.0413$

dove $W \sim \chi^2(3)$ cioè ha la stessa densità di T sotto $\mathbb{H}_0$ (asintoticamente).

Quindi, per es., rifiuto $\mathbb{H}_0$ se $\alpha = 0.05$, accetto $\mathbb{H}_0$ se $\alpha = 0.025$.

Test per l'indipendenza

Supponiamo che ogni elemento di una popolazione possa essere classificato in base a due caratteristiche o criteri che denotiamo con X e Y.

X ha r valori possibili: $1, \ldots, r$.
Y ha s valori possibili: $1, \ldots, s$.

Seleziono a caso n individui nella popolazione e considero le corrispondenti caratteristiche (X_i, Y_i) $i = 1, \ldots, n$

Posso assumere i vettori (X_i, Y_i) $i = 1, \ldots, n$ indipendenti, mentre in genere le caratteristiche X_i e Y_i non sono indipendenti.

Anzi il nostro obiettivo è verificare se lo siano oppure no.

$(X_1, Y_1), \ldots, (X_n, Y_n)$ campione aleatorio

$(X_i, Y_i) \sim (X, Y)$ $i = 1, \ldots, n$ e $i.i.d.$

Denotiamo con $p_{i,j} := P(X = i, Y = j)$, $p_i := P(X = i)$, $q_j := P(Y = j)$.

Sulla base del campione vogliamo verificare

$\mathbb{H}_0 : p_{i,j} = p_i q_j$ per $i = 1, \ldots, r$ e $j = 1, \ldots, s$
contro l'ipotesi alternativa
$\mathbb{H}_1 : p_{i,j} \neq p_i q_j$ per qualche $i = 1, \ldots, r$ e $j = 1, \ldots, s$.

Poniamo per $i = 1, \ldots, r$ e $j = 1, \ldots, s$

- $N_{i,j} = n^o$ degli elementi del campione che sono uguali a (i, j); (frequenza di (i, j) nel campione)
- $N_i = \sum_{j=1}^{s} N_{i,j} = n^o$ delle $X_1, \ldots, X_n$ che sono uguali a i; (frequenza di i nel campione $X_1, \ldots, X_n$)
- $M_j = \sum_{i=1}^{r} N_{i,j} = n^o$ delle $Y_1, \ldots, Y_n$ che sono uguali a j; (frequenza di j nel campione $Y_1, \ldots, Y_n$)
- $\hat{p}_i = \dfrac{N_i}{n}$ e $\hat{q}_j = \dfrac{M_j}{n}$.

Si considera la statistica-test:

$$T^* := \sum_{j=1}^{s} \sum_{i=1}^{r} \frac{(N_{i,j} - n\hat{p}_i\hat{q}_j)^2}{n\hat{p}_i\hat{q}_j}.$$

Si può dimostrare che se "n è grande" e sotto $\mathbb{H}_0$

$$T^* \approx \chi^2((r-1)\times(s-1)) \;\Rightarrow\; P_{\mathbb{H}_0}\left(T^* \geq \chi^2_{\alpha,(r-1)\times(s-1)}\right) \simeq \alpha.$$

Idea del test

Per ogni (i,j), $N_{i,j} = n^o$ degli elementi del campione che sono uguali a (i,j)

$\Downarrow$

$N_{i,j} \sim Bin(n, p_{i,j} = P(X = i, Y = j))$

$\Downarrow$

$\mathbb{E}(N_{i,j}) = np_{i,j} \overset{\text{sotto } \mathbb{H}_0}{=} np_i q_j = nP(X = i)P(Y = j)$

$\Downarrow$

$(N_{i,j} - np_i q_j)^2$ grande sta ad indicare che l'ipotesi nulla $\mathbb{H}_0$ di indipendenza non è plausibile e poiché p_i e q_j sono incognite le stimo con $\hat{p}_i = \frac{N_i}{n}$ e $\hat{q}_j = \frac{M_j}{n}$

$\Downarrow$

statistica-test $T^* := \sum_{j=1}^{s} \sum_{i=1}^{r} \frac{(N_{i,j} - n\hat{p}_i \hat{q}_j)^2}{n\hat{p}_i \hat{q}_j}$

$\Downarrow$

rifiuto $\mathbb{H}_0$ se $T^* \geq c$, la soglia c è scelta in base al livello di significatività α.

Formula utile per la statistica-test

$$T^* = \sum_{j=1}^{s}\sum_{i=1}^{r} \frac{(N_{i,j})^2}{n\hat{p}_i\hat{q}_j} - n = \sum_{j=1}^{s}\sum_{i=1}^{r} \frac{(N_{i,j})^2}{\frac{N_i M_j}{n}} - n.$$

Quindi un test approssimato di livello α per verificare:

$\mathbb{H}_0$: le caratteristiche X e Y sono indipendenti i.e.

$p_{i,j} = p_i q_j$ per $i = 1, \ldots, r$ e $j = 1, \ldots, s$

contro l'ipotesi alternativa

$\mathbb{H}_1$: le caratteristiche X e Y non sono indipendenti i.e.

$p_{i,j} \neq p_i q_j$ per qualche $i = 1, \ldots, r$ e $j = 1, \ldots, s$

è quello che, avendo osservato il valore $T^* = t^*$

$$\text{rifiuto } \mathbb{H}_0 \text{ se } t^* \geq \chi^2_{\alpha,(r-1)\times(s-1)}$$

$$\text{accetto } \mathbb{H}_0 \text{ se } t^* < \chi^2_{\alpha,(r-1)\times(s-1)}$$

$p - value = P_{\mathbb{H}_0}(T^* \geq t^*) \simeq P(W^* \geq t^*)$ con $W^* \sim \chi^2((r-1)\times(s-1))$.

Esempio 1. Un certo corso universitario viene impartito a studenti del secondo anno di 3 diversi indirizzi; gli studenti frequentano le lezioni del medesimo professore che registra il numero di studenti di ogni indirizzo che hanno superato l'esame nell'arco dell'anno. I dati sono i seguenti

X/Y	ind. 1	ind. 2	ind. 3	Tot. esami	$\hat{p}_\cdot = \frac{N_\cdot}{n}$
esame superato	$N_{1,1} = 30$	$N_{1,2} = 15$	$N_{1,3} = 50$	$N_1 = 95$	$\hat{p}_1 = \frac{95}{180}$
esame non superato	$N_{2,1} = 40$	$N_{2,2} = 8$	$N_{2,3} = 37$	$N_2 = 85$	$\hat{p}_2 = \frac{85}{180}$
tot. stud. ind. i	$M_1 = 70$	$M_2 = 23$	$M_3 = 87$	$n = 180$	
$\hat{q}_\cdot = \frac{M_\cdot}{n}$	$\hat{q}_1 = \frac{70}{180}$	$\hat{q}_2 = \frac{23}{180}$	$\hat{q}_3 = \frac{87}{180}$		

Il rendimento degli studenti dei 3 indirizzi, relativamente all'esame in questione, si può ritenere sostanzialmente equivalente oppure le differenze sono sistematicamente significative?

Quanto chiesto sopra equivale a chiedersi se la variabile
$X =$ "risultato dell'esame"
e la variabile
$Y =$ "indirizzo"
sono indipendenti o no.
valori di $X \Rightarrow \{1,2\}$ $r = 2$
valori di $Y \Rightarrow \{1,2,3\}$ $s = 3$

$T^* := \sum_{j=1}^{3}\sum_{i=1}^{2} \frac{N_{i,j}^2}{\frac{N_i M_j}{n}} - n \simeq 4.9554$ è il valore della statistica-test

$T^* \overset{\mathbb{H}_0}{\approx} \chi^2\big((r-1)\times(s-1)\big) = \chi^2\big((2-1)\times(3-1)\big) = \chi^2(2)$

$p - value = P_{\mathbb{H}_0}(T^* \geq 4.9554) \simeq P(W^* \geq 4.9554) \simeq 0.0837$

dove $W^* \sim \chi^2(2) \Rightarrow$ si accetta al 5% con i dati a disposizione.

Esempio 2. I gruppi sanguigni di un campione di persone sono classificati in base al loro tipo e al fattore Rh. I dati raccolti sono riassunti dalla tabella:

Rh / tipo	0	A	B	AB
positivo	72	74	18	6
negativo	12	14	2	2

Si può affermare che tipo e fattore Rh siano caratteristiche indipendenti del gruppo sanguigno?

Soluzione

X=Rh / Y=tipo	0	A	B	AB	N_i
positivo	72	74	18	6	170
negativo	12	14	2	2	30
M_j	84	88	20	8	$n = 200$

$$T^* = \sum_{i=1}^{2}\sum_{j=1}^{4} \frac{N_{i,j}^2}{\frac{N_i M_j}{n}} - n$$

$$= \frac{(72)^2}{\frac{84\times 170}{200}} + \frac{(74)^2}{\frac{88\times 170}{200}} + \frac{(18)^2}{\frac{20\times 170}{200}} + \frac{6^2}{\frac{8\times 170}{200}} +$$

$$+ \frac{(12)^2}{\frac{84\times 30}{200}} + \frac{(14)^2}{\frac{88\times 30}{200}} + \frac{2^2}{\frac{20\times 30}{200}} + \frac{2^2}{\frac{8\times 30}{200}} - 200 \simeq 1.11$$

$n=200$, $(r-1)\times(s-1)=3$ e $\chi^2_{0.05,3} \simeq 7.815 \Rightarrow$ non rifiuto $\mathbb{H}_0$ al 5%.

Esempio 3. Sono state effettuate delle prove di resistenza su pneumatici di 4 diverse marche, e si è registrata la durata di questi pneumatici (in km percorsi prima dell'usura). I dati sono i seguenti

$N_{i,j}$	Marca A $j=1$	Marca B $j=2$	Marca C $j=3$	Marca D $j=4$	N_i
i=1 durata $< 30000km$	26	23	15	32	96
i=2 durata tra $30000e45000km$	118	93	116	121	448
i=3 durata $> 45000km$	56	84	69	47	256
$M.$	200	200	200	200	$n = 800$

Ci chiediamo se le 4 marche si possono ritenere equivalenti, quanto alla durata dei pneumatici oppure no.

In altre parole questo equivale a chiedersi se la variabile "durata" sia indipendente dalla variabile "marca'.

N.B. Indipendenza tra le variabili mi dice che conoscere la marca non altera la valutazione della probabilità che lo pneumatico duri di più o di meno.

Soluzione. La statistica-test è:

$$T^* = \sum_{i=1}^{3}\sum_{j=1}^{4} \frac{N_{i,j}^2}{\frac{N_i M_j}{n}} - n \simeq 20.7723$$

$N_1 = 96 \qquad N_2 = 448 \qquad N_3 = 256$

$M_1 = 200 \qquad M_2 = 200 \qquad M_3 = 200 \qquad M_4 = 200$

sotto $\mathbb{H}_0$ e per "n grande":
$T^* \approx \chi^2\big((4-1)\times(3-1) = 6)\big) \qquad \chi^2_{0.05,6} = 12.592$

$\Rightarrow$ rifiuto $\mathbb{H}_0$ al 5%.

$p - value = P_{\mathbb{H}_0}(T^* \geq 20.7723) \simeq 0.002$

quindi l'ipotesi di indipendenza può essere rifiutata a qualsiasi livello $\geq 0.2\%$.