

9. Test del χ^2 e test di Smirnov-Kolmogorov

9.1 Stimatori di massima verosimiglianza per distribuzioni con densità finita

Supponiamo di avere un campione statistico X_1, \dots, X_n e di sapere che esso è relativo ad una distribuzione su un insieme finito t_1, \dots, t_k . Dunque conosco la distribuzione se conosco $p_j := \mathbb{P}(X_i = t_j)$ per ogni $j = 1, \dots, k$.

Dato il campione sperimentale x_1, \dots, x_n , cerchiamo gli stimatori di massima verosimiglianza per i parametri p_1, \dots, p_k . Tra i dati rilevati x_1, \dots, x_n ce ne sono:

n_1 che valgono t_1 ,

n_2 che valgono t_2 ,

\dots ,

n_k che valgono t_k ,

con la condizione $n_1 + n_2 + \dots + n_k = n$.

La densità congiunta di (X_1, \dots, X_n) in x_1, \dots, x_n è dunque

$$g(x_1, \dots, x_n | p_1, \dots, p_k) = p_1^{n_1} p_2^{n_2} \dots p_k^{n_k} = \prod_{j=1}^k p_j^{n_j}$$

e perciò

$$h(x_1, \dots, x_n | p_1, \dots, p_k) := \log g(x_1, \dots, x_n | p_1, \dots, p_k) = \sum_{j=1}^k n_j \log p_j.$$

Usiamo i moltiplicatori di Lagrange per massimizzare g rispetto ai p_1, \dots, p_k ammissibili:

$$H(p_1, \dots, p_k, \lambda) = \sum_{j=1}^k n_j \log p_j - \lambda \left(\sum_{j=1}^k p_j - 1 \right).$$

$$\frac{\partial h}{\partial \lambda} = - \left(\sum_{j=1}^k p_j - 1 \right), \quad \frac{\partial h}{\partial p_j} = \frac{n_j}{p_j} - \lambda \quad \forall j = 1, \dots, k.$$

Da cui otteniamo

$$p_j = \frac{n_j}{n} \quad \forall j = 1, \dots, k,$$

ovvero lo stimatore di massima verosimiglianza per la densità in t_j è la frequenza relativa del carattere t_j :

$$P_j = \frac{1}{n} \# \{i \in \{1, \dots, n\} : X_i = t_j\}.$$

9.2 Esperimenti binari e distribuzione binomiale

Supponiamo di effettuare un esperimento che può avere solo due risultati (che indichiamo come *successo* e *insuccesso*). Ad ogni ripetizione dell'esperimento si può ottenere *successo* con probabilità $p \in [0, 1]$ e *insuccesso* con probabilità $1 - p$.

Ripetiamo l'esperimento n volte, sempre nelle stesse condizioni, in modo che le varie ripetizioni dell'esperimento siano tra loro indipendenti.

Otteniamo una stringa ordinata di lunghezza n (x_1, x_2, \dots, x_n) dove x_i vale successo o insuccesso. Se la stringa contiene k successi e $n - k$ insuccessi, la probabilità di ottenere esattamente quella stringa ordinata è $p^k(1 - p)^{n-k}$. Quante sono le stringhe contenenti esattamente k successi e $n - k$ insuccessi? Poiché ogni stringa è individuata dalla posizione dei suoi *successi*, esse sono tante quanti sono i sottoinsiemi di k elementi che si possono ottenere da un insieme di n elementi, ovvero $\binom{n}{k}$. Dunque la probabilità di ottenere k *successi* e $n - k$ *insuccessi* è $\binom{n}{k}p^k(1 - p)^{n-k}$. In altre parole, la v.a. che *conta i successi* ha distribuzione binomiale di parametri n e p .

Per $i = 1, \dots, n$, indico con X_i la v.a. che vale 1 se all' i -esima ripetizione dell'esperimento ottengo un *successo* e vale 0 se ottengo *insuccesso*. Le v.a. X_1, \dots, X_n sono indipendenti e identicamente distribuite con $\mathbb{P}_{X_i} = B(p)$. Osserviamo che la v.a. che *conta i successi* non è altro che la somma delle X_1, \dots, X_n dunque: la somma di n v.a. i.i.d. con distribuzione $B(p)$ è una variabile aleatoria con distribuzione binomiale $B(n, p)$.

9.3 Test del χ^2

Sia Y_1, \dots, Y_n un campione statistico. Supponiamo che le v.a. del campione siano discrete a valori t_1, \dots, t_k . La distribuzione è nota se conosco la densità di probabilità in ciascun punto $t_j, j = 1, \dots, k$

$$p_j := \mathbb{P}(Y_i = t_j), \quad j = 1, \dots, k.$$

Siano p_1^0, \dots, p_k^0 dei numeri assegnati, tali che $p_j^0 \geq 0 \quad \forall j = 1, \dots, k$ e $\sum_{j=1}^k p_j^0 = 1$.

Vogliamo testare

$$H_0: p_j = p_j^0 \quad \forall j = 1, \dots, k \quad H_A: \exists \bar{j} \in \{1, \dots, k\}: p_{\bar{j}} \neq p_{\bar{j}}^0.$$

Per ogni $j = 1, \dots, k$ considero

$$X_j = \# \{i \in \{1, \dots, n\}: Y_i = t_j\} \quad j = 1, \dots, k.$$

Se chiamo *successo* l'evento *ottengo il valore t_j* , vediamo subito che X_j è la v.a. che conta i successi in n ripetizioni indipendenti di un esperimento in cui la probabilità di successo in ogni singola prova è p_j , dunque

$$\mathbb{P}_{X_j} = B(n, p_j), \quad \mathbb{E}[X_j] = np_j, \quad \text{Var}[X_j] = np_j(1 - p_j).$$

Inoltre $(X_j - np_j^0)^2$ mi dice quanto sia verosimile che $\mathbb{P}(Y_i = t_j) = p_j$ sia uguale a p_j^0 . Posso stabilire un criterio di accettazione considerando una opportuna combinazione lineare $\sum_{j=1}^k a_j (X_j - np_j^0)^2$ con coefficienti a_1, \dots, a_k positivi. Si può dimostrare che vale il seguente

Teorema 9.3.1 (di Pearson). Se $P_{X_j} = \text{Bin}(n, p_j^0)$, allora la v.a. $\sum_{j=1}^k \frac{(X_j - np_j^0)^2}{np_j^0}$, per $n \rightarrow \infty$, converge in legge ad una v.a. con distribuzione χ_{k-1}^2 .

Osservazione 9.3.1. L'approssimazione è considerata accettabile se $np_j^0 \geq 5 \quad \forall j = 1, \dots, k$.

Formuliamo allora il seguente criterio di accettazione. Siano n_1, \dots, n_k le frequenze assolute (o effettivi) dei caratteri t_1, \dots, t_k nel campione empirico x_1, \dots, x_n

$$\text{accetto } H_0 \text{ se } t_n := \sum_{j=1}^k \frac{(n_j - np_j^0)^2}{np_j^0} < \varepsilon. \text{ Rifiuto altrimenti}$$

La probabilità di commettere errore di prima specie è allora

$$\alpha := \mathbb{P} \left(\sum_{j=1}^k \frac{(X_j - np_j^0)^2}{np_j^0} \geq \varepsilon \mid p_j = p_j^0 \quad \forall j = 1, \dots, k \right) \simeq 1 - F_{\chi_{k-1}^2}(\varepsilon).$$

Scelgo dunque ε tale che $F_{\chi_{k-1}^2}(\varepsilon) = 1 - \alpha$, cioè $\varepsilon = \chi_{k-1, 1-\alpha}^2$. Dunque:

$$\text{Accetto } H_0 \text{ se e solo se } \sum_{j=1}^k \frac{(n_j - np_j^0)^2}{np_j^0} < \chi_{k-1, 1-\alpha}^2$$

Osservazione 9.3.2. Non dimostriamo il Teorema 9.3.1 ma ne vediamo la sua *plausibilità* nel caso $k = 2$, cioè il campione Y_1, \dots, Y_n può assumere solo i valori t_1 e t_2 .

Considero $Z_i := \mathbb{1}_{\{Y_i=t_1\}}$. Allora Z_1, \dots, Z_n sono i.i.d. con $\mathbb{P}(Z_1) = \text{Ber}(p_1^0)$ e $X_1 = \sum_{i=1}^n Z_i$. Si ha inoltre

$$T = \frac{(X_1 - np_1^0)^2}{np_1^0} + \frac{(X_2 - np_2^0)^2}{np_2^0}, \quad p_1^0 + p_2^0 = 1, \quad X_1 + X_2 = n,$$

da cui

$$T = \frac{(X_1 - np_1^0)^2}{np_1^0} + \frac{(X_1 - np_1^0)^2}{n(1 - p_1^0)} = \frac{(X_1 - np_1^0)^2}{np_1^0(1 - p_1^0)} = \left(\frac{\sum_{i=1}^n Z_i - n\mathbb{E}[Z_1]}{\sqrt{n\text{Var}[Z_1]}} \right)^2.$$

Per il teorema del limite centrale $\frac{\sum_{i=1}^n Z_i - n\mathbb{E}[Z_1]}{\sqrt{n\text{Var}[Z_1]}}$ converge in legge a una v.a. gaussiana standard e sappiamo che il quadrato di una v.a. con distribuzione $N(0, 1)$ segue la distribuzione χ^2 ad un grado di libertà.

9.4 Test di Kolmogorov-Smirnov

È un test sulla legge del campione. Vediamo un risultato preliminare

Lemma 9.4.1. Se X è una v.a. con legge F , allora $F(X)$ è uniformemente distribuita sull'intervallo $[0, 1]$.

Dimostrazione. Dimostriamo il lemma limitatamente al caso assolutamente continuo. Sia f la densità della distribuzione di X : $\mathbb{P}_X = f(x)dx$ e sia $\psi: \mathbb{R} \rightarrow \mathbb{R}$ una funzione di Borel non negativa. Si ha

$$\int_{\mathbb{R}} \psi(t) \mathbb{P}_{F(X)} dt = \int_{\mathbb{R}} \psi(F(x)) \mathbb{P}_X(dx) = \int_{\mathbb{R}} \psi(F(x)) f(x) dx = \int_0^1 \psi(t) dt$$

dove abbiamo effettuato il cambio di variabile $t = F(x)$. □

Teorema 9.4.2. *Sia X_1, \dots, X_n campione statistico con legge continua F . Per $i = 1, \dots, n$ e $t \in \mathbb{R}$ pongo*

$$Y_i(\omega, t) = \mathbb{1}_{(-\infty, t]}(X_i(\omega)) = \begin{cases} 1 & X_i(\omega) \leq t, \\ 0 & X_i(\omega) > t. \end{cases}$$

Allora, per ogni $t \in \mathbb{R}$, le v.a. $Y_1(\cdot, t), \dots, Y_n(\cdot, t)$ sono un campione statistico con distribuzione di Bernoulli di parametro $p = F(t)$. In particolare Si ha

$$\mathbb{E}[Y_i(\cdot, t)] = \mathbb{P}(X_i \leq t) = F_0(t), \quad \text{Var}[Y_i(\cdot, t)] = F_0(t)(1 - F_0(t)) \leq \frac{1}{4}.$$

Sia inoltre

$$G_n(\omega, t) = \frac{1}{n} \sum_{i=1}^n Y_i(\omega, t) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{(-\infty, t]}(X_i(\omega)).$$

e sia

$$D_n(\omega) := \sup_{t \in \mathbb{R}} |G_n(\omega, t) - F(t)|.$$

Allora

1. $\mathbb{P}(|G_n(\cdot, t) - F_0(t)| > \varepsilon) \leq \frac{1}{4n\varepsilon^2} \quad \forall \varepsilon > 0, \quad \forall t \in \mathbb{R}.$
2. La legge di D_n non dipende da F .

Dimostrazione. Per dimostrare il primo punto è sufficiente osservare che

$$\mathbb{E}[G_n(\cdot, t)] = F(t), \quad \text{Var}[G_n(\cdot, t)] = \frac{1}{n} F(t)(1 - F(t)) \leq \frac{1}{4n}.$$

Vediamo il secondo punto: sia $d \geq 0$. Calcolo

$$\begin{aligned} \mathbb{P}(D_n \geq d) &= \mathbb{P}\left(\sup_{t \in \mathbb{R}} \left| \frac{1}{n} \#\{i: X_i \leq t\} - F(t) \right| \geq d\right) = \\ &= \mathbb{P}\left(\sup_{t \in \mathbb{R}} \left| \frac{1}{n} \#\{i: F(X_i) \leq F(t)\} - F(t) \right| \geq d\right). \end{aligned}$$

Infatti, se F è strettamente crescente, allora $X_i \leq t$ se e solo se $F(X_i) \leq F(t)$. Se invece F è crescente, ma non strettamente, l'uguaglianza rimane vera a livello di probabilità perché la probabilità che X_i cada in un intervallo in cui F è costante è comunque nulla.

D'altra parte le v.a. $U_i := F(X_i)$ sono i.i.d. con distribuzione uniforme sull'intervallo $[0, 1]$, dunque

$$\begin{aligned} \mathbb{P}(D_n \geq d) &= \mathbb{P}\left(\sup_{t \in \mathbb{R}} \left| \frac{1}{n} \#\{i: U_i \leq F(t)\} - F(t) \right| \geq d\right) = \\ &= \mathbb{P}\left(\sup_{y \in (0,1)} \left| \frac{1}{n} \#\{i: U_i \leq y\} - y \right| \geq d\right) \end{aligned}$$

dato che, essendo continua, F assume tutti i valori compresi tra il suo estremo inferiore ed il suo estremo superiore. \square

Osservazione 9.4.1. Osserviamo che $G_n(\omega, t) = \frac{1}{n} \#\{i \in \{1, \dots, n\}: X_i(\omega) \leq t\}$ dunque $G_n(\omega, \cdot)$ è una funzione costante a tratti, monotona crescente che prende valori in $0, \frac{1}{n}, \dots, \frac{n-1}{n}, 1$ (li prende tutti se e solo se i valori $X_1(\omega), \dots, X_n(\omega)$ sono tutti distinti).

Osservazione 9.4.2. Si può dimostrare che vale il seguente limite

$$\lim_{n \rightarrow \infty} \mathbb{P}(D_n \sqrt{n} \leq t) = \begin{cases} 1 - 2 \sum_{j=1}^{\infty} (-1)^{j-1} \exp(-2j^2 t^2) & t > 0, \\ 0 & t \leq 0. \end{cases}$$

Test di Kolmogorov-Smirnov Consideriamo il seguente test d'ipotesi per un campione statistico X_1, \dots, X_n di cui rilevo i dati x_1, \dots, x_n . Sia $F_0: \mathbb{R} \rightarrow [0, 1]$ una funzione continua monotona crescente, tale che $\lim_{t \rightarrow -\infty} F_0(t) = 0, \lim_{t \rightarrow +\infty} F_0(t) = 1$. Voglio testare

$$H_0: F_0 \text{ è la legge del campione,} \quad H_A: \exists t \in \mathbb{R}: F_0(t) \neq \mathbb{P}(X_i \leq t).$$

Sia $d_n := \sup_{t \in \mathbb{R}} |g_n(x_1, \dots, x_n, t) - F_0(t)|$. Accetto H_0 se $d_n < \varepsilon$, rifiuto altrimenti.

Vediamo se possiamo scegliere ε in base al livello di significatività desiderato. Riconsideriamo la probabilità di commettere errore di prima specie.

$$\mathbb{P}(D_n \geq \varepsilon) = \mathbb{P}(D_n \sqrt{n} \geq \varepsilon \sqrt{n}) \sim 2 \sum_{j=1}^{\infty} (-1)^{j-1} \exp(-2j^2 \varepsilon^2 n) \leq 2 \exp(-2\varepsilon^2 n).$$

Scegliamo dunque $\varepsilon > 0$ tale che $\alpha = 2 \exp(-2\varepsilon^2 n)$ cioè $\varepsilon = \sqrt{\frac{1}{2n} \log \frac{2}{\alpha}}$. Quindi

$$\text{accetto } H_0 \text{ se e solo se } \sup_{t \in \mathbb{R}} \left| \frac{1}{n} \#\{i: x_i \leq t\} - F_0(t) \right| < \sqrt{\frac{1}{2n} \log \frac{2}{\alpha}}.$$

Osservazione 9.4.3. Supponiamo di aver ordinato i dati x_1, \dots, x_n in ordine crescente (per

semplicità supponiamo che siano tutti distinti). Abbiamo

$$\begin{aligned}
 \sup_{t \in \mathbb{R}} \left| \frac{1}{n} \# \{i: x_i \leq t\} - F_0(t) \right| &= \max \left\{ \sup_{t < x_1} \left| \frac{1}{n} \# \{i: x_i \leq t\} - F_0(t) \right|, \right. \\
 &\quad \sup_{t \in [x_1, x_2)} \left| \frac{1}{n} \# \{i: x_i \leq t\} - F_0(t) \right|, \dots, \sup_{t \in [x_{n-1}, x_n)} \left| \frac{1}{n} \# \{i: x_i \leq t\} - F_0(t) \right|, \\
 &\quad \left. \sup_{t \geq x_n} \left| \frac{1}{n} \# \{i: x_i \leq t\} - F_0(t) \right| \right\} \\
 &= \max \left\{ \sup_{t < x_1} |F_0(t)|, \sup_{t \in [x_1, x_2)} \left| \frac{1}{n} - F_0(t) \right|, \dots, \sup_{t \in [x_{n-1}, x_n)} \left| \frac{n-1}{n} - F_0(t) \right|, \sup_{t \geq x_n} |1 - F_0(t)| \right\} \\
 &= \max \left\{ F(x_1), \left| \frac{1}{n} - F(x_1) \right|, \left| \frac{1}{n} - F(x_2) \right|, \dots, \right. \\
 &\quad \left. \left| \frac{n-1}{n} - F(x_{n-1}) \right|, \left| \frac{n-1}{n} - F(x_n) \right|, |1 - F(x_n)| \right\}.
 \end{aligned}$$