Inferenza statistica

Def Un campione statistico è un insieme $X_1, \dots, X_n$ di v.a. indipendenti ed equidistribuite, indicando con F la loro funzione di ripartizione.
La numerosità n del campione si chiama taglia del campione.
Indichiamo con $\mu = E[X_i] \ \forall i = 1, \dots, n$,
$\sigma^2 = Var(X_i) \ \forall i = 1, \dots, n$.

Si chiamano dati un insieme $x_1, \dots, x_n$ di numeri reali, con l'idea che $x_i = X_i(\omega)$ $\forall i = 1, \dots, n$ e per un $\omega \in \Omega$.

Def Si chiama statistica campionaria una funzione $g(X_1, \dots, X_n)$. Si dice che una statistica campionaria è corretta se $E[g(X_1, \dots, X_n)] =$ parametro da misurare.

ES Media campionaria

$$\overline{X} := \frac{X_1 + \dots + X_n}{n}$$

OSS $\overline{X}(\omega) = \frac{1}{n} \sum_{i=1}^{n} x_i = \overline{x}$ media empirica

$\overline{X}$ è una stima corretta del valore atteso

$$E[\overline{X}] = E\left[\frac{X_1 + \dots + X_n}{n}\right] = \mu$$

ES <u>Varianza campionaria</u>

$$S^2 := \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2$$

Prop S^2 è una stima corretta della varianza σ^2.

dim
$$\sum_{i=1}^{n} (X_i - \overline{X})^2 = \sum_{i=1}^{n} X_i^2 - n\overline{X}^2$$

$$\left[\sum_{i=1}^{n} (X_i - \overline{X})^2 = \sum_{i=1}^{n} X_i^2 + n\overline{X}^2 - 2\overline{X}\sum_{i=1}^{n} X_i = \right.$$
$$\left. = \sum_{i=1}^{n} X_i^2 + n\overline{X}^2 - 2\overline{X}(n\overline{X}) = \sum_{i=1}^{n} X_i^2 - n\overline{X}^2 \right]$$

$$E\left[\sum_{i=1}^{n} (X_i - \overline{X})^2\right] = E\left[\sum_{i=1}^{n} X_i^2\right] - nE[\overline{X}^2] =$$
$$= \sum_{i=1}^{n} E[X_i^2] - nE[\overline{X}^2] =$$

$$\sigma^2 = Var(X_i) = E[X_i^2] - \mu^2, \quad Var(\overline{X}) = E[\overline{X}^2] - \mu^2$$
$$Var(\overline{X}) = Var\left(\frac{X_1 + \dots + X_n}{n}\right) = \frac{\sigma^2}{n}$$

$$= \sum_{i=1}^{n} (\sigma^2 + \mu^2) - n\left(\frac{\sigma^2}{n} + \mu^2\right) =$$
$$= n\sigma^2 + n\mu^2 - \sigma^2 - n\mu^2 = (n-1)\sigma^2$$
$$\Longrightarrow E[S^2] = \sigma^2 \qquad \square$$

OSS $S^2(\omega) = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$ varianza empirica

---

Supponiamo che la legge del campione dipenda da un parametro $\vartheta \in A \subseteq \mathbb{R}$. Vogliamo dare una stima di ϑ dipendente dai dati.

Stimatori puntuali

- Massimo verosimiglianza $\longrightarrow \hat{\vartheta}$
- Metodo dei momenti $\longrightarrow \tilde{\vartheta}$

① Massima verosimiglianza

$\{X_i\}$ abbiano funzione di massa p_ϑ o densità f_ϑ.

Costruiamo la funzione

$$A \ni \vartheta \longmapsto L(\vartheta; x_1, \dots, x_n) := \begin{cases} \prod_{i=1}^{n} p_\vartheta(x_i) & \text{(discreto)} \\ \prod_{i=1}^{n} f_\vartheta(x_i) & \text{(con densità)} \end{cases}$$

e cerchiamo il punto di massimo $\hat{\vartheta}$.

② Metodo dei momenti

Cerchiamo $\tilde{\vartheta}$ per cui $E_\vartheta[X^k] = \frac{1}{n} \sum_{i=1}^{n} x_i^k$, $k \geq 1 \dots$

Esempi $\quad X_k \sim B(2,p)$, $\quad p \in (0,1)$

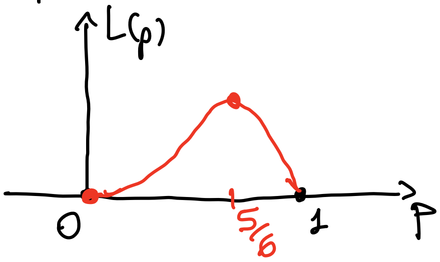
① $(0,1) \ni p \longmapsto L(p; x_1,\dots,x_n) = \prod_{i=1}^{n} p_p(x_i)$

$$L(p; x_1,\dots,x_n) = \prod_{i=1}^{n} \binom{2}{x_i} p^{x_i} (1-p)^{2-x_i} =$$

$$= p^{x_1+\dots+x_n} (1-p)^{2n-(x_1+\dots+x_n)} \prod_{i=1}^{n} \binom{2}{x_i}$$

ES $\quad n=3$, $\quad x_1 = 2$, $\quad x_2 = 1$, $\quad x_3 = 2$.

$$L(p) = L(p; 2,1,2) = p^5 (1-p)^{6-5} \prod_{i=1}^{3} \binom{2}{x_i}$$



$$L(p) = p^5 (1-p) \prod_{i=1}^{3} \binom{2}{x_i}, \quad L'(p) = 5p^4(1-p)\prod\binom{2}{x_i} - p^5 \prod\binom{2}{x_i}$$

$$= p^4 \prod\binom{2}{x_i}\left[5(1-p)-p\right] =$$

$$= p^4 \prod\binom{2}{x_i}\left[5-6p\right]$$

$$L'(p) = 0 \iff p = \frac{5}{6}, \quad p = 0$$

$$\hat{p} = \frac{5}{6}$$

CRITERIO $\quad$ Se $L(\vartheta) > 0 \ \forall \vartheta \in A$ e L è derivabile in A,

$$L'(\vartheta) = 0 \iff \frac{d}{d\vartheta}\left(\log L(\vartheta)\right) = 0 \qquad \left[\frac{d}{d\vartheta} \log L(\vartheta) = \frac{L'(\vartheta)}{L(\vartheta)}\right]$$

$$L(p; x_1, \dots, x_n) = p^{x_1+\dots+x_n} (1-p)^{2n-(x_1+\dots+x_n)} \prod_{i=1}^{n} \binom{2}{x_i}$$

$$L'(p) = 0 \iff \frac{d}{dp}\left[(x_1+\dots+x_n)\log p + (2n-(x_1+\dots+x_n))\log(1-p) + \log \prod_{i=1}^{n}\binom{2}{x_i} \right] = 0$$

$$\iff \frac{x_1+\dots+x_n}{p} + \frac{2n-(x_1+\dots+x_n)}{1-p}(-1) = 0$$

$$\iff p = \frac{1}{2n}\sum_{i=1}^{n} x_i = \frac{\overline{x}}{2} \implies \boxed{\hat{p} = \frac{\overline{x}}{2}}$$

<u>OSS</u> Se $X_1, \dots, X_n$ v.a. indip. e $B(2,p)$, abbiamo costruito la statistica campionaria

$$p(X_1, \dots, X_n) = \frac{\overline{X}}{2} = \frac{1}{2n}\sum_{i=1}^{n} X_i$$

che è la stima del parametro p data dal metodo di massima verosimiglianza.

$$E[p(X_1, \dots, X_n)] = E\left[\frac{\overline{X}}{2}\right] = \frac{\mu}{2} = \frac{2p}{2} = p$$

ed è una stima corretta.

<u>Prop</u> Se $X_1, \dots, X_n \sim B(k,p)$, una stima corretta del parametro p è $p(X_1, \dots, X_n) = \frac{1}{k}\overline{X}$.

<u>OSS</u> $\hat{p} = \frac{1}{k}\overline{x}$

(2) $E_p[X] = \frac{1}{n}\sum_{i=1}^{n} x_i = \overline{x}$ $\qquad X_k \sim B(2,p)$

<u>esempio</u> $n=3$, $x_1 = 2$, $x_2 = 1$, $x_3 = 2$

$$\boxed{\begin{array}{l} E_p[X] = 2p \quad , \quad \bar{x} = \frac{1}{3}\cdot 5 = \frac{5}{3} \\ \tilde{p} \text{ è soluzione di } \quad 2p = \frac{5}{3} \Rightarrow \tilde{p} = \frac{5}{6} \end{array}}$$

$$E_p[X] = 2p \quad , \quad \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$\tilde{p}$ è soluzione di $\quad 2p = \bar{x} \Rightarrow \boxed{\tilde{p} = \frac{\bar{x}}{2}}$

OSS $\quad X_i \sim B(k,p)$ allora

$$\hat{p} = \tilde{p} = \frac{\bar{x}}{k}$$

ES $\quad X_1,\dots,X_n$ gaussiane $\mathcal{N}(\mu,\sigma^2)$

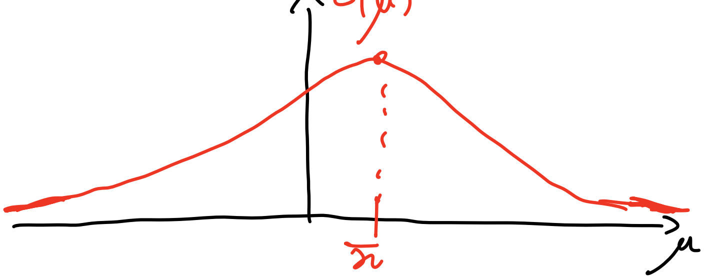
Supponiamo che σ sia nota e stimiamo $\mu \in \mathbb{R}$.

(1)

$$\mathbb{R} \ni \mu \longmapsto L(\mu; x_1,\dots,x_n) = \prod_{i=1}^{n} f_\mu(x_i) =$$

$$= \prod_{i=1}^{n} \left[\frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} \right]$$

$$L(\mu; x_1,\dots,x_n) = \left(\frac{1}{\sqrt{2\pi}\,\sigma}\right)^n e^{-\frac{1}{2\sigma^2}\left[\sum_{i=1}^{n}(x_i-\mu)^2\right]}$$



$$L'(\mu)=0 \iff \frac{d}{d\mu}\left[n \log \frac{1}{\sqrt{2\pi}\,\sigma} - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2 \right] = 0$$

$$\iff -\frac{1}{2\sigma^2}\sum_{i=1}^{n} 2(x_i-\mu)(-1) = 0$$

$$\iff \sum_{i=1}^{n}(x_i-\mu) = 0 \iff \mu = \frac{1}{n}\sum_{i=1}^{n} x_i = \overline{x}$$

$$\Rightarrow \boxed{\hat{\mu} = \overline{x}}$$

ES $\quad \hat{\sigma^2} = ? \qquad \hat{\sigma^2}(X_1, \dots X_n) = ?$

② $\quad E_\mu[X] = \frac{1}{n}\sum_{i=1}^{n} x_i \quad , \quad X \sim N(\mu, \sigma^2)$

$$\Rightarrow \boxed{\tilde{\mu} = \overline{x}}$$

Esperimento: Lanciamo tre volte un insieme di monete oneste molto numeroso. Otteniamo 100 teste, 90 teste e 110 teste. Stimare la numerosità?

$$n = 3 \ , \quad X_1, X_2, X_3 \qquad X \sim B(k, \tfrac{1}{2})$$

① $\cancel{\mathbb{N}}$ $(0,+\infty)$ $\ni k \longmapsto L(k; x_1, x_2, x_3) = \prod_{i=1}^{3} \binom{k}{x_i} \left(\frac{1}{2}\right)^{x_i} \left(\frac{1}{2}\right)^{k-x_i}$

$$L(k) = \left(\frac{1}{2}\right)^{3k} \prod_{i=1}^{3} \binom{k}{x_i}$$

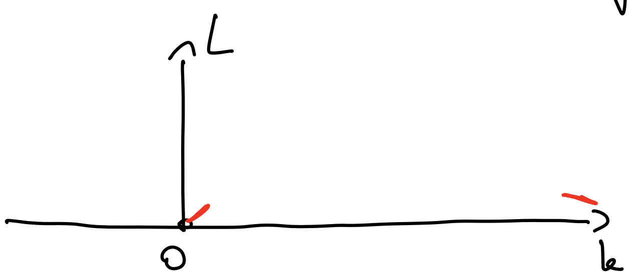
Teo del limite centrale $\Longrightarrow B(k, \tfrac{1}{2}) \underset{k \text{ grande}}{\approx} N\left(\underset{\mu}{\tfrac{k}{2}}, \underset{\sigma^2}{\tfrac{k}{4}}\right)$

$$\frac{B(k,\tfrac{1}{2}) - k \cdot \tfrac{1}{2}}{\sqrt{k \cdot \tfrac{1}{4}}} \approx N(0,1)$$

$(0,+\infty) \ni k$

$$L(k; x_1, x_2, x_3) = \prod_{i=1}^{3} \frac{1}{\sqrt{2\pi \frac{k}{4}}} \cdot e^{-\frac{(x_i - k/2)^2}{2\, k/4}} =$$

$$= \left(\sqrt{\frac{2}{\pi}}\sqrt{\frac{1}{k}}\right)^3 e^{-\frac{2}{k}\sum_{i=1}^{3}\left(x_i - \frac{k}{2}\right)^2}$$



$$L'(k) = 0 \iff \frac{d}{dk}\left[\log\left(\sqrt{\frac{2}{\pi}}\right)^3 + \log k^{-\frac{3}{2}} - \frac{2}{k}\sum_{i=1}^{3}\left(x_i - \frac{k}{2}\right)^2\right] = 0$$

$$\iff -\frac{3}{2k} + \frac{2}{k^2}\sum_{i=1}^{3}\left(x_i - \frac{k}{2}\right)^2 - \frac{2}{k}\sum_{i=1}^{3} 2\left(x_i - \frac{k}{2}\right)\left(-\frac{1}{2}\right) = 0$$

$$\iff \frac{1}{k^2}\left[-\frac{3}{2}k + 2\sum_{i=1}^{3}\left(x_i - \frac{k}{2}\right)^2 + 2k\sum_{i=1}^{3}\left(x_i - \frac{k}{2}\right)\right] = 0$$

$x_1 = 100, \quad x_2 = 90, \quad x_3 = 110 \quad \Longrightarrow \boxed{\hat{k} \sim 200.332}$

(2) $\quad E_k[X] = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad X \sim B(k, \tfrac{1}{2})$

$\frac{k}{2} = \bar{x} \quad \Longrightarrow \boxed{\tilde{k} = 2\bar{x}}$

Esperimento $\quad n = 3, \; x_1 = 100, \; x_2 = 90, \; x_3 = 110$

$\Longrightarrow \tilde{k} = 200$