5. Stimatori di massima versosimiglianza

Sia $X_1, \ldots, X_n$ un campione statistico e sia $Y = \varphi(X_1, \ldots, X_n)$ una sua statistica. Se Y ha lo scopo di stimare un parametro θ della distribuzione del campione, diciamo che Y è uno *stimatore del parametro* θ.

Supponiamo di conoscere la distribuzione del campione a meno di un parametro θ e supponiamo che tale distribuzione sia discreta o assolutamente continua e dunque dotata di densità (discreta o meno). Tale densità dipenderà dal parametro θ e la indico col simbolo $g(x|\theta)$. La densità congiunta si indica col simbolo $f(x_1, \ldots, x_n|\theta)$ e sappiamo che, grazie all'indipendenza delle v.a. che costituiscono il campione, si ha

$$f(x_1, \ldots, x_n|\theta) = g(x_1|\theta) \cdot \ldots \cdot g(x_n|\theta) = \prod_{i=1}^{n} g(x_i|\theta).$$

Interpreto $f(x_1, \ldots, x_n|\theta)$ come la *plausibilità* che la n-upla $x_1, \ldots, x_n$ si realizzi nel campione empirico quando il parametro incognito prende il valore θ. Consideriamo infatti i due casi, discreto e assolutamente continuo

- Se il campione ha distribuzione discreta, allora

$$\mathbb{P}(X_1 = x_1, \ldots, X_n = x_n) = \prod_{i=1}^{n} \mathbb{P}(X_i = x_i) = \prod_{i=1}^{n} g(x_i|\theta).$$

- Se il campione ha distribuzione assolutamente continua, e se la densità g è continua, allora, per $\delta > 0$ e sufficientemente piccolo si ha

$$\begin{aligned}\mathbb{P}\left(|X_1 - x_1| < \frac{\delta}{2}, \ldots, |X_n - x_n| < \frac{\delta}{2}\right) &= \prod_{i=1}^{n} \mathbb{P}\left(|X_i - x_i| < \frac{\delta}{2}\right) \\ &= \prod_{i=1}^{n} \mathbb{P}\left(X_i \in \left(x_i - \frac{\delta}{2}, x_i + \frac{\delta}{2}\right)\right) \simeq \prod_{i=1}^{n} \left(g\left(x_i|\theta\right)\delta\right)^n \\ &= \delta^n \prod_{i=1}^{n} g\left(x_i|\theta\right) = \delta^n f(x_1, \ldots, x_n|\theta)\end{aligned}$$

La funzione f, vista come funzione di θ, si dice *funzione di versimiglianza.*

Dunque: dato il campione empirico $x_1, \ldots, x_n$, cerco $\widehat{\theta} = \widehat{\theta}(x_1, \ldots, x_n)$ che massimizza la funzione $f(x_1, \ldots, x_n|\theta)$. La statistica $\widehat{\theta} = \widehat{\theta}(X_1, \ldots, X_n)$ si dirà *stimatore di massima verosimiglianza del parametro* θ.

Osservazione 5.0.1. Poiché la funzione $\ln\colon (0,+\infty) \to \mathbb{R}$ è strettamente monotona crescente, massimizzare $f(x_1,\ \ldots,\ x_n|\theta) = \prod_{i=1}^n g(x_i|\theta)$ equivale a massimizzare la funzione

$$h(x_1,\ \ldots,\ x_n|\theta) := \ln f(x_1,\ \ldots,\ x_n|\theta) = \sum_{i=1}^n \ln g(x_i|\theta)$$

Si ha

$$\frac{\partial}{\partial\theta} h(x_1,\ \ldots,\ x_n|\theta) = \frac{\partial}{\partial\theta} \sum_{i=1}^n \ln g(x_i|\theta) = \sum_{i=1}^n \frac{\partial}{\partial\theta} \ln g(x_i|\theta) = \sum_{i=1}^n \frac{1}{g(x_i|\theta)} \frac{\partial g(x_i|\theta)}{\partial\theta}$$

5.1 Distribuzione di Bernoulli

In questo caso la distribuzione dipende dal solo parametro $p = \mathbb{P}(X_i = 1) = \mathbb{E}[X_i]$.
Sia dunque $X_1,\ \ldots,\ X_n$ un campione statistico di Bernoulli di parametro incognito $p \in [0,1]$. Realizzo n prove di Bernoulli e ottengo il campione empirico $x_1,\ \ldots,\ x_n$, $x_i \in \{0,1\}$. Sia $k = k(x_1,\ \ldots,\ x_n) := \sum_{i=1}^n x_i$. Abbiamo

$$\begin{aligned}
f(x_1,\ \ldots,\ x_n|p) &= \mathbb{P}(X_1 = x_1, \ldots, X_n = x_n) = p^k(1-p)^{n-k},\\
h(x_1,\ \ldots,\ x_n|p) &= \ln(p^k(1-p)^{n-k}) = k\ln p + (n-k)\ln(1-p).\\
\frac{\partial h}{\partial p} &= \frac{k}{p} - \frac{n-k}{1-p} = \frac{k-np}{p(1-p)} \gtreqless 0 \iff k - np \gtreqless 0 \iff p \lesseqgtr \frac{k}{n}.
\end{aligned}$$

Poiché $k = \sum_{i=1}^n x_i$, lo stimatore di massima verosimiglianza per il parametro p è $\frac{\sum_{i=1}^n X_i}{n}$ cioè la media campionaria $\overline{X}$. Ricordiamo che la media campionaria è uno stimatore corretto di $\mathbb{E}[X_i] = p$.

5.2 Distribuzione di Poisson

La distribuzione di Poisson è concentrata sugli interi nonnegativi e dipende da un solo parametro:

$$g(x|\lambda) = \mathrm{e}^{-\lambda}\,\frac{\lambda^x}{x!}, \qquad x = 0, 1, \ldots.$$

Dunque

$$f(x_1,\ \ldots,\ x_n|\lambda) = \prod_{i=1}^n \left(\mathrm{e}^{-\lambda}\,\frac{\lambda^{x_i}}{x_i!}\right),$$

$$\begin{aligned}
h(x_1,\ \ldots,\ x_n|\lambda) = \ln f(x_1,\ \ldots,\ x_n|\lambda) &= \sum_{i=1}^n \ln\left(\mathrm{e}^{-\lambda}\,\frac{\lambda^{x_i}}{x_i!}\right)\\
&= \sum_{i=1}^n (-\lambda + x_i\ln(\lambda) - \ln(x_i!)) = -n\lambda + n\overline{x}\ln(\lambda) - \sum_{i=1}^n \ln(x_i!)
\end{aligned}$$

Da cui

$$\frac{\partial}{\partial\lambda} h(x_1,\ \ldots,\ x_n|\lambda) = n\left(-1 + \frac{\overline{x}}{\lambda}\right) \gtreqless 0 \iff \lambda \lesseqgtr \overline{x}.$$

Quindi anche in questo caso o stimatore di massima verosimiglianza per il parametro λ è la media campionaria $\overline{X}$ (che è uno stimatore corretto).

5.3 Distribuzione gaussiana

In questo caso la densità dipende da due parametri, $\mu \in \mathbb{R}$ e $\sigma > 0$:

$$\begin{aligned} f(x_1,\ \ldots,\ x_n|\mu,\sigma) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(x_i-\mu)^2}{2\sigma^2}\right) \\ &= (2\pi)^{\frac{-n}{2}} (\sigma)^{-n} \exp\left(\frac{-\sum_{i=1}^{n}(x_i-\mu)^2}{2\sigma^2}\right) \end{aligned}$$

cosicché

$$\begin{aligned} h(x_1,\ \ldots,\ x_n|\mu,\sigma) = \ln f(x_1,\ \ldots,\ x_n|\mu,\sigma) &= -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2 \\ &= -\frac{n}{2}\ln(2\pi) - n\ln(\sigma) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2. \end{aligned}$$

Si ha quindi

$$\begin{aligned} \frac{\partial}{\partial\mu} h(x_1,\ \ldots,\ x_n|\mu,\sigma) &= \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu) = n\left(\overline{x}-\mu\right), \\ \frac{\partial}{\partial\sigma} h(x_1,\ \ldots,\ x_n|\mu,\sigma) &= \frac{-n}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^{n}(x_i-\mu)^2 = \frac{1}{\sigma^3}\left(-n\sigma^2 + \sum_{i=1}^{n}(x_i-\mu)^2\right). \end{aligned}$$

Dunque le due derivate parziali si annullano contemporaneamente se e solo se

$$\mu = \overline{x}, \qquad \sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^2 = \frac{n-1}{n}\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\overline{x})^2 = \frac{n-1}{n}s^2.$$

Dunque la media campionaria $\overline{X}$ è uno stimatore di massima verosimiglianza (ed è uno stimatore corretto) per il valore atteso μ mentre $\dfrac{n-1}{n}S^2$ è uno stimatore di massima verosimiglianza per la varianza σ^2.

5.4 Distribuzione uniforme su un intervallo

Se (a,b) è l'intervallo, allora la densità del campione è

$$g(x|a,b) = \begin{cases} \frac{1}{b-a} & x \in [a,b], \\ 0 & \text{altrimenti} \end{cases}$$

da cui

$$f(x_1,\ \ldots,\ x_n|a,b) = \begin{cases} \frac{1}{(b-a)^n} & x_i \in [a,b] \quad \forall i = 1,\ldots,n, \\ 0 & \text{altrimenti.} \end{cases}$$

Devo massimizzare $\frac{1}{(b-a)^n}$ con il vincolo $a \le x_i \le b$ per ogni $i = 1,\ldots,n$. Devo dunque minimizzare la lunghezza dell'intervallo $b-a$ con il vincolo $a \le x_i \le b$ per ogni $i = 1,\ldots,n$. È dunque

$$a = \min\{x_1,\ \ldots,\ x_n\}, \qquad b = \max\{x_1,\ \ldots,\ x_n\}.$$

Dunque

$$\min\{X_1,\ \ldots,\ X_n\}, \qquad \max\{X_1,\ \ldots,\ X_n\}$$

sono stimatori di massima verosimiglianza rispettivamente per l'estremo inferiore e per l'estremo superiore dell'intervallo.

5.5 Efficienza degli stimatori

Sia $X_1,\ \ldots,\ X_n$ un campione statistico. Supponiamo di conoscere la *forma* della distribuzione ma che essa dipenda da parametro incognito θ (scalare o vettoriale).

Supponiamo che $T = t(X_1,\ \ldots,\ X_n)$ sia uno stimatore del parametro (scalare) θ. T potrebbe non essere uno stimatore corretto. Definisco *bias dello stimatore θ la quantità*

$$b_\theta(T) := \mathbb{E}[T] - \theta.$$

Valuto l'efficacia di T considerando l'errore quadratico medio

$$\begin{aligned} r(T,\theta) :=& \mathbb{E}\left[(T-\theta)^2\right] = \mathbb{E}\left[(T - \mathbb{E}[T] + b_\theta(T))^2\right] \\ =& \mathbb{E}\left[(T-\mathbb{E}[T])^2 + 2b_\theta(T)(T-\mathbb{E}[T]) + b_\theta^2(T)\right] = \operatorname{Var}[T] + b_\theta^2(T) \end{aligned}$$

Supponiamo che $X_1,\ \ldots,\ X_n$ sia un campione statistico con distribuzione uniforme sull'intervallo $[0,\theta]$. Sappiamo che

$$\mathbb{E}[X_i] = \frac{\theta}{2}, \qquad \operatorname{Var}[X_i] = \frac{\theta^2}{12}.$$

Come stimatore del parametro θ considero allora la statistica $T = 2\overline{X} = \dfrac{2}{n}\displaystyle\sum_{i=1}^{n} X_i$. Abbiamo

$$\mathbb{E}[T] = \theta, \qquad b_\theta(T) = 0, \qquad r(T,\theta) = \operatorname{Var}[T] = 4\operatorname{Var}\left[\overline{X}\right] = \frac{\theta^2}{3n}.$$

Considero ora lo stimatore di massima verosimiglianza

$$\widehat{T} := \max(X_1,\ \ldots,\ X_n).$$

Per calcolarne valore atteso e varianza ne calcoliamo innanzitutto la legge:

$$F_{\widehat{T}}(t) = \mathbb{P}(\widehat{T} \le t) = \mathbb{P}(X_1 \le t, \ldots, X_n \le t) = \prod_{i=1}^{n} \mathbb{P}(X_i \le t) = \begin{cases} 0 & t < 0, \\ \left(\dfrac{t}{\theta}\right)^n & 0 \le t < \theta, \\ 1 & t \ge \theta. \end{cases}$$

Dunque $\mathbb{P}_{\widehat{T}} = g(t)\mathrm{d}t$ con

$$g(t) = \begin{cases} n\dfrac{t^{n-1}}{\theta^n}, & t \in [0,\theta], \\ 0 & \text{altrimenti} \end{cases}$$

Abbiamo dunque

$$\begin{aligned}
\mathbb{E}\left[\widehat{T}\right] &= \int_{\mathbb{R}} t g(t)\mathrm{d}t = \int_0^\theta \frac{n}{\theta^n} t^n \mathrm{d}t = \frac{n\,\theta}{n+1},\\
\mathbb{E}\left[\widehat{T}^2\right] &= \int_{\mathbb{R}} t^2 g(t)\mathrm{d}t = \int_0^\theta \frac{n}{\theta^n} t^{n+1} \mathrm{d}t = \frac{n\,\theta^2}{n+2},\\
\mathrm{Var}\left[\widehat{T}\right] &= \frac{n\,\theta^2}{n+2} - \left(\frac{n\,\theta}{n+1}\right)^2 = \frac{n\,\theta^2}{(n+2)(n+1)^2}.
\end{aligned}$$

Quindi

$$\begin{aligned}
b_\theta(\widehat{T}) :=& \mathbb{E}\left[\widehat{T}\right] - \theta = \frac{n\,\theta}{n+1} - \theta = \frac{-\,\theta}{n+1},\\
r(\widehat{T},\theta) =& \mathbb{E}\left[(\widehat{T}-\theta)^2\right] = \mathrm{Var}\left[\widehat{T}\right] + b_\theta^2(\widehat{T}) = \frac{n\,\theta^2}{(n+2)(n+1)^2} + \frac{\theta^2}{(n+1)^2}\\
&= \frac{2\,\theta^2}{(n+1)(n+2)}
\end{aligned}$$

Dunque, per n grande $r(\widehat{T},\theta) << r(T,\theta)$ anche se $\mathbb{E}\left[\widehat{T}\right] = \dfrac{n\,\theta}{n+1} \neq \theta = \mathbb{E}\left[T\right]$ e dunque T è uno stimatore corretto mentre $\widehat{T}$ non lo è.

Cerchiamo ora il *miglior stimatore* nella famiglia $T_c := c\widehat{T}$, $c \in \mathbb{R}$ dove per *miglior stimatore* intendiamo quello che minimizza l'errore quadratico medio $r(T_c,\theta)$. Abbiamo

$$\mathbb{E}\left[T_c\right] = c\mathbb{E}\left[\widehat{T}\right] = \frac{c\,n\,\theta}{n+1}, \qquad \mathrm{Var}\left[T_c\right] = c^2\mathrm{Var}\left[\widehat{T}\right] = \frac{c^2\,n\,\theta^2}{(n+2)(n+1)^2}.$$

Quindi

$$\begin{aligned}
b_\theta(T_c) =& \mathbb{E}\left[T_c\right] - \theta = \frac{c\,n\,\theta}{n+1} - \theta = \frac{\theta(nc-n-1)}{n+1},\\
r(T_c,\theta) =& \mathrm{Var}\left[T_b\right] + b_\theta^2(T_c) = \frac{c^2\,n\,\theta^2}{(n+2)(n+1)^2} + \frac{\theta^2(nc-n-1)^2}{(n+1)^2}.
\end{aligned}$$

Abbiamo allora

$$\begin{aligned}
&\frac{\partial}{\partial\theta} r(T_c,\theta) = \frac{2\,c\,n\,\theta^2}{(n+2)(n+1)^2} + \frac{2n\,\theta^2(nc-n-1)}{(n+1)^2}\\
&= \frac{2n\,\theta^2}{(n+1)^2}\left[\frac{c}{n+2} + (nc-n-1)\right] = \frac{2n\,\theta^2}{(n+2)(n+1)^2}\left[c(n+1)^2 - (n+2)(n+1)\right]\\
&= \frac{2n\,\theta^2}{(n+2)(n+1)}\left[c(n+1) - (n+2)\right] \gtreqless 0 \iff c \gtreqless \frac{n+2}{n+1}
\end{aligned}$$

Quindi $c \in \mathbb{R} \mapsto r(T_c,\theta)$ ammette minimo assoluto nel punto $c := \dfrac{n+2}{n+1}$ ed il minimo vale

$$r(T_{\frac{n+2}{n+1}},\theta) = \mathrm{Var}\left[T_b\right] + b_\theta^2(T_c) = \frac{\left(\frac{n+2}{n+1}\right)^2 n\,\theta^2}{(n+2)(n+1)^2} + \frac{\theta^2(n\frac{n+2}{n+1}-n-1)^2}{(n+1)^2} = \frac{\theta^2}{(n+1)^2}$$

6. Intervalli di confidenza

La media campionaria e la varianza campionaria ci offrono una stima dei parametri valore atteso e varianza del campione statistico in esame. Abbiamo però bisogno di sapere *quanto ci si possa fidare di questa stima* ovvero quale sia la probabilità che il *vero* valore del parametro incognito non sia *troppo distante* dalla stima trovata.

Diamo perciò la seguente definizione:

Definizione 6.0.1 (Intervallo di confidenza)**.** Sia $X_1, \ldots, X_n$ un campione statistico e sia θ un parametro (ignoto) che caratterizza la distribuzione del campione.
Siano $L_i = l_i(X_1, \ldots, X_n)$ e $L_s = l_s(X_1, \ldots, X_n)$ due statistiche del campione e sia $\alpha \in (0,1)$. Dico che l'intervallo (L_i, L_s) è un *intervallo di confidenza* (o di fiducia) di livello $1-\alpha$ se $\mathbb{P}(\theta \in (L_i, L_s)) \geq 1-\alpha$, ovvero che (L_i, L_s) è un intervallo di confidenza (o di fiducia) di errore α se $\mathbb{P}(\theta \notin (L_i, L_s)) \leq \alpha$.

Dico che la semiretta $(L_i, +\infty)$ è un *intervallo di confidenza unilaterale superiore* di livello $1-\alpha$ se $\mathbb{P}(\theta > L_i) \geq 1-\alpha$

Dico che la semiretta $(-\infty, L_s)$ è un *intervallo di confidenza unilaterale inferiore* di livello $1-\alpha$ se $\mathbb{P}(\theta < L_s) \geq 1-\alpha$

Osservazione 6.0.1.

1. La scelta dei nomi delle due statistiche non è casuale: L_i sta per limitazione inferiore mentre L_s sta per limitazione superiore.
2. Di solito si è interessati a *piccoli* valori di α, più precisamente a $\alpha \in (10^{-2}, 10^{-1})$.
3. La disuguaglianza di Chebychev ci ha fornito un intervallo di confidenza per il valore atteso μ del campione nel caso in cui la varianza σ^2 sia nota

$$\mathbb{P}\left(\left|\overline{X} - \mu\right| \geq t\right) \leq \frac{\sigma^2}{t^2} \qquad \forall t > 0$$

ovvero

$$\mathbb{P}\left(\left|\overline{X} - \mu\right| < t\right) \geq 1 - \frac{\sigma^2}{t^2} \qquad \forall t > 0$$

cioè

$$\mathbb{P}\left(\overline{X} - t < \mu < \overline{X} + t\right) \geq 1 - \frac{\sigma^2}{t^2} \qquad \forall t > 0.$$

Fissato $\alpha \in (0,1)$ scelgo $t = \dfrac{\sigma}{\sqrt{\alpha}}$. La disuguaglianza di Chebychev si legge allora

$$\mathbb{P}\left(\overline{X} - \frac{\sigma}{\sqrt{\alpha}} < \mu < \overline{X} + \frac{\sigma}{\sqrt{\alpha}}\right) \geq 1 - \alpha \qquad \forall \alpha \in (0,1).$$

Dunque l'intervallo $\left(\overline{X} - \dfrac{\sigma}{\sqrt{\alpha}}, \overline{X} + \dfrac{\sigma}{\sqrt{\alpha}}\right)$ è un intervallo di confidenza di livello $1-\alpha$ per il valore atteso μ del campione.

6.1 Stima per intervalli del valore atteso di campioni gaussiani

6.1.1 Campione gaussiano di cui è nota la varianza

Intervallo bilaterale

Sia $X_1, \ldots, X_n$ un campione gaussiano di valore atteso μ incognita e varianza σ^2 nota.

Sia Z una v.a. gaussiana standard e sia $\alpha \in (0,1)$. Calcolo $\mathbb{P}\left(|Z| \leq z_{1-\frac{\alpha}{2}}\right)$:

$$\begin{aligned}
&\mathbb{P}\left(|Z| \leq z_{1-\frac{\alpha}{2}}\right) = \mathbb{P}\left(-z_{1-\frac{\alpha}{2}} \leq Z \leq z_{1-\frac{\alpha}{2}}\right) \\
&= \mathbb{P}\left(Z \leq z_{1-\frac{\alpha}{2}}\right) - \mathbb{P}\left(Z \leq -z_{1-\frac{\alpha}{2}}\right) = \mathbb{P}\left(Z \leq z_{1-\frac{\alpha}{2}}\right) - \mathbb{P}\left(Z \leq z_{\frac{\alpha}{2}}\right) \qquad (6.1) \\
&= \Phi\left(z_{1-\frac{\alpha}{2}}\right) - \Phi\left(z_{\frac{\alpha}{2}}\right) = 1 - \frac{\alpha}{2} - \frac{\alpha}{2} = 1 - \alpha.
\end{aligned}$$

Sappiamo che $\mathbb{P}_{\overline{X}} = N\left(\mu, \frac{\sigma^2}{n}\right)$ e che dunque $\dfrac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$ ha distribuzione $N(0,1)$. Applichiamo quindi la disuguaglianza (6.1) a $\dfrac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$. Si ha:

$$\begin{aligned}
1 - \alpha &= \mathbb{P}\left(\frac{\left|\overline{X} - \mu\right|}{\frac{\sigma}{\sqrt{n}}} \leq z_{1-\frac{\alpha}{2}}\right) = \mathbb{P}\left(-z_{1-\frac{\alpha}{2}} \leq \frac{\mu - \overline{X}}{\frac{\sigma}{\sqrt{n}}} \leq z_{1-\frac{\alpha}{2}}\right) \\
&= \mathbb{P}\left(\frac{-\sigma\, z_{1-\frac{\alpha}{2}}}{\sqrt{n}} \leq \mu - \overline{X} \leq \frac{\sigma\, z_{1-\frac{\alpha}{2}}}{\sqrt{n}}\right) \\
&= \mathbb{P}\left(\overline{X} - \frac{\sigma\, z_{1-\frac{\alpha}{2}}}{\sqrt{n}} \leq \mu \leq \overline{X} + \frac{\sigma\, z_{1-\frac{\alpha}{2}}}{\sqrt{n}}\right)
\end{aligned}$$

L'intervallo

$$\left(\overline{X} - \frac{\sigma\, z_{1-\frac{\alpha}{2}}}{\sqrt{n}}, \overline{X} + \frac{\sigma\, z_{1-\frac{\alpha}{2}}}{\sqrt{n}}\right)$$

è dunque un intervallo di confidenza di livello $1-\alpha$ per il valore atteso μ del campione.

Osservazione 6.1.1 (Dimensionamento del campione)**.** Fissato il livello di confidenza $1-\alpha$, supponiamo di voler controllare l'ampiezza dell'intervallo di confidenza $L_s - L_i$. Nel caso in esame l'ampiezza dell'intervallo di confidenza è $\dfrac{2\,\sigma\, z_{1-\frac{\alpha}{2}}}{\sqrt{n}}$. Se fissiamo una limitazione superiore $2\,\delta$ per l'ampiezza di tale intervallo, deve dunque essere

$$\frac{2\,\sigma\, z_{1-\frac{\alpha}{2}}}{\sqrt{n}} \leq 2\,\delta$$

ovvero

$$n \geq \left(\frac{\sigma\, z_{1-\frac{\alpha}{2}}}{\delta}\right)^2.$$

Intervallo unilaterale superiore

Sia Z una v.a. tale che $\mathbb{P}_Z = N(0,1)$. Sappiamo che

$$\mathbb{P}(Z \leq t) = 1 - \alpha \quad \text{ se e solo se } \quad t = z_{1-\alpha}.$$

Abbiamo dunque

$$1-\alpha = \mathbb{P}\left(\frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \le z_{1-\alpha}\right) = \mathbb{P}\left(\overline{X}-\mu \le \frac{\sigma\, z_{1-\alpha}}{\sqrt{n}}\right) = \mathbb{P}\left(\mu \ge \overline{X} - \frac{\sigma\, z_{1-\alpha}}{\sqrt{n}}\right).$$

Quindi la semiretta

$$\left(\overline{X} - \frac{\sigma\, z_{1-\alpha}}{\sqrt{n}}, +\infty\right)$$

è un intervallo di confidenza unilaterale superiore di livello $1-\alpha$.

Intervallo unilaterale inferiore

Sia Z una v.a. tale che $\mathbb{P}_Z = N(0,1)$. Sappiamo che

$$\mathbb{P}(Z \ge t) = 1-\alpha \quad \text{se e solo se} \quad \mathbb{P}(Z \le t) = \alpha \quad \text{se e solo se} \quad t = z_\alpha.$$

Abbiamo dunque

$$1-\alpha = \mathbb{P}\left(\frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \ge z_{\alpha}\right) = \mathbb{P}\left(\overline{X}-\mu \ge \frac{\sigma\, z_{\alpha}}{\sqrt{n}}\right) = \mathbb{P}\left(\mu \le \overline{X} - \frac{\sigma\, z_{\alpha}}{\sqrt{n}}\right).$$

Quindi la semiretta

$$\left(-\infty, \overline{X} - \frac{\sigma\, z_{\alpha}}{\sqrt{n}}\right) = \left(-\infty, \overline{X} + \frac{\sigma\, z_{1-\alpha}}{\sqrt{n}}\right)$$

è un intervallo di confidenza unilaterale inferiore di livello $1-\alpha$.

6.1.2 Campione gaussiano di cui non è nota la varianza

Intervallo bilaterale

Sia $X_1, \ldots, X_n$ un campione gaussiano di valore atteso μ varianza σ^2, entrambe incognite.

Sappiamo che la v.a. $T := \dfrac{(\overline{X}-\mu)\sqrt{n}}{S}$ segue la distribuzione t di Student con $n-1$ gradi di libertà:

$$\mathbb{P}_T = t(n-1).$$

Sia $t_{n-1,1-\frac{\alpha}{2}}$ il relativo quantile di livello $1-\frac{\alpha}{2}$:

$$\mathbb{P}\left(T \le t_{n-1,1-\frac{\alpha}{2}}\right) = 1 - \frac{\alpha}{2}.$$

Calcolo $\mathbb{P}\left(|T| \le t_{n-1,1-\frac{\alpha}{2}}\right)$:

$$\begin{aligned}
\mathbb{P}\left(|T| \le t_{n-1,1-\frac{\alpha}{2}}\right) &= \mathbb{P}\left(-t_{n-1,1-\frac{\alpha}{2}} \le T \le t_{n-1,1-\frac{\alpha}{2}}\right) \\
&= \mathbb{P}\left(T \le t_{n-1,1-\frac{\alpha}{2}}\right) - \mathbb{P}\left(T \le -t_{n-1,1-\frac{\alpha}{2}}\right) \\
&= \mathbb{P}\left(T \le t_{n-1,1-\frac{\alpha}{2}}\right) - \mathbb{P}\left(T \le t_{n-1,\frac{\alpha}{2}}\right) = 1 - \frac{\alpha}{2} - \frac{\alpha}{2} = 1-\alpha.
\end{aligned}$$

Abbiamo dunque

$$\begin{aligned}1-\alpha &= \mathbb{P}\left(|T| \le t_{n-1,1-\frac{\alpha}{2}}\right) = \mathbb{P}\left(\frac{\left|\overline{X}-\mu\right|\sqrt{n}}{S} \le t_{n-1,1-\frac{\alpha}{2}}\right)\\ &= \mathbb{P}\left(\left|\overline{X}-\mu\right| \le \frac{S}{\sqrt{n}} t_{n-1,1-\frac{\alpha}{2}}\right)\\ &= \mathbb{P}\left(\frac{-S}{\sqrt{n}} t_{n-1,1-\frac{\alpha}{2}} \le \mu - \overline{X} \le \frac{S}{\sqrt{n}} t_{n-1,1-\frac{\alpha}{2}}\right)\\ &= \mathbb{P}\left(\overline{X} - \frac{S}{\sqrt{n}} t_{n-1,1-\frac{\alpha}{2}} \le \mu \le \overline{X} + \frac{S}{\sqrt{n}} t_{n-1,1-\frac{\alpha}{2}}\right)\end{aligned}$$

L'intervallo

$$\left(\overline{X} - \frac{S\, t_{n-1,1-\frac{\alpha}{2}}}{\sqrt{n}}, \overline{X} + \frac{S\, t_{n-1,1-\frac{\alpha}{2}}}{\sqrt{n}}\right)$$

è dunque un intervallo di confidenza di livello $1-\alpha$ per il valore atteso μ del campione.

Intervallo unilaterale superiore

Sappiamo che

$$\mathbb{P}\left(T \le t\right) = 1-\alpha \quad \text{ se e solo se } \quad t = t_{n-1,1-\alpha}.$$

Abbiamo dunque

$$\begin{aligned}1-\alpha &= \mathbb{P}\left(\frac{(\overline{X}-\mu)\sqrt{n}}{S} \le t_{n-1,1-\alpha}\right) = \mathbb{P}\left(\overline{X}-\mu \le \frac{S\, t_{n-1,1-\alpha}}{\sqrt{n}}\right)\\ &= \mathbb{P}\left(\mu \ge \overline{X} - \frac{S\, t_{n-1,1-\alpha}}{\sqrt{n}}\right).\end{aligned}$$

Quindi la semiretta

$$\left(\overline{X} - \frac{S\, t_{n-1,1-\alpha}}{\sqrt{n}}, +\infty\right)$$

è un intervallo di confidenza unilaterale superiore di livello $1-\alpha$.

Intervallo unilaterale inferiore

Sappiamo che

$$\mathbb{P}\left(T \ge t\right) = 1-\alpha \quad \text{ se e solo se } \quad \mathbb{P}\left(T \le t\right) = \alpha \quad \text{ se e solo se } \quad t = t_{n-1,\alpha}.$$

Abbiamo dunque

$$1-\alpha = \mathbb{P}\left(\frac{(\overline{X}-\mu)\sqrt{n}}{S} \ge t_{n-1,\alpha}\right) = \mathbb{P}\left(\overline{X}-\mu \ge \frac{S\, t_{n-1,\alpha}}{\sqrt{n}}\right) = \mathbb{P}\left(\mu \le \overline{X} - \frac{S\, t_{n-1,\alpha}}{\sqrt{n}}\right).$$

Quindi la semiretta

$$\left(-\infty, \overline{X} - \frac{S\, t_{n-1,\alpha}}{\sqrt{n}}\right) = \left(-\infty, \overline{X} + \frac{S\, t_{n-1,1-\alpha}}{\sqrt{n}}\right)$$

è un intervallo di confidenza unilaterale inferiore di livello $1-\alpha$.

6.2 Stima per intervalli della varianza di campioni gaussiani

Intervallo bilaterale

Sia $X_1, \ldots, X_n$ un campione gaussiano di valore atteso μ (incognita o nota)e varianza σ^2 incognita.

Sappiamo che la v.a. $V := (n-1)\dfrac{S^2}{\sigma^2}$ segue la distribuzione χ^2 a $n-1$ gradi di libertà. Per ogni $\alpha \in (0,1)$ indico con $\chi^2_{n-1,\alpha}$ il quantile di livello α della v.a. V:

$$F_V\left(\chi^2_{n-1,\alpha}\right) = \alpha \qquad \forall \alpha \in (0,1).$$

Osservazione 6.2.1. $\chi^2_{n-1,\alpha} > 0$ per ogni $\alpha \in (0,1)$.

Calcolo $\mathbb{P}\left(\chi^2_{n-1,\frac{\alpha}{2}} < V < \chi^2_{n-1,1-\frac{\alpha}{2}}\right)$:

$$\mathbb{P}\left(\chi^2_{n-1,\frac{\alpha}{2}} < V < \chi^2_{n-1,1-\frac{\alpha}{2}}\right) = \mathbb{P}\left(V < \chi^2_{n-1,1-\frac{\alpha}{2}}\right) - \\ - \mathbb{P}\left(V < \chi^2_{n-1,\frac{\alpha}{2}}\right) = 1 - \frac{\alpha}{2} - \frac{\alpha}{2} = 1-\alpha.$$

Dunque

$$1-\alpha = \mathbb{P}\left(\chi^2_{n-1,\frac{\alpha}{2}} < (n-1)\frac{S^2}{\sigma^2} < \chi^2_{n-1,1-\frac{\alpha}{2}}\right)$$
$$= \mathbb{P}\left(\frac{1}{\chi^2_{n-1,1-\frac{\alpha}{2}}} < \frac{\sigma^2}{(n-1)S^2} < \frac{1}{\chi^2_{n-1,\frac{\alpha}{2}}}\right) = \mathbb{P}\left(\frac{(n-1)S^2}{\chi^2_{n-1,1-\frac{\alpha}{2}}} < \sigma^2 < \frac{(n-1)S^2}{\chi^2_{n-1,\frac{\alpha}{2}}}\right)$$

Quindi l'intervallo

$$\left(\frac{(n-1)S^2}{\chi^2_{n-1,1-\frac{\alpha}{2}}}, \frac{(n-1)S^2}{\chi^2_{n-1,\frac{\alpha}{2}}}\right)$$

è un intervallo di confidenza di livello $1-\alpha$ per la varianza σ^2 del campione.

Intervallo unilaterale superiore

Sappiamo che

$$\mathbb{P}\left(V \le t\right) = 1-\alpha \quad \text{ se e solo se } \quad t = \chi^2_{n-1,1-\alpha}.$$

Dunque

$$1-\alpha = \mathbb{P}\left((n-1)\frac{S^2}{\sigma^2} < \chi^2_{n-1,1-\alpha}\right) = \mathbb{P}\left(\sigma^2 > (n-1)\frac{S^2}{\chi^2_{n-1,1-\alpha}}\right).$$

Quindi la semiretta

$$\left(\frac{(n-1)S^2}{\chi^2_{n-1,1-\alpha}}, +\infty\right)$$

è un intervallo di confidenza di livello $1-\alpha$ per la varianza σ^2 del campione.

Intervallo unilaterale inferiore

Sappiamo che

$$\mathbb{P}(V \geq t) = 1 - \alpha \quad \text{se e solo se} \quad \mathbb{P}(V \leq t) = \alpha \quad \text{se e solo se} \quad t = \chi^2_{n-1,\alpha}.$$

Dunque

$$1 - \alpha = \mathbb{P}\left((n-1)\frac{S^2}{\sigma^2} > \chi^2_{n-1,\alpha}\right) = \mathbb{P}\left(\sigma^2 \leq (n-1)\frac{S^2}{\chi^2_{n-1,\alpha}}\right).$$

Quindi l'intervallo

$$\left(0, \frac{(n-1)S^2}{\chi^2_{n-1,\alpha}}\right)$$

è un intervallo di confidenza di livello $1 - \alpha$ per la varianza σ^2 del campione.

Esempio 6.2.1. Calcoliamo gli intervalli di confidenza per il carattere Totpor dei dati tratti da [2], nell'ipotesi che si tratti della realizzazione di v.a. normali.

```
> setwd("~/Documents/didattica/2017-18_analisi_reale/alcuni_appunti/esempio_statistica")
>
> library(readr)
>
> table2 <- read_delim("~/Documents/didattica/2017-18_analisi_reale/alcuni_appunti/
table2.csv", "\t", escape_double = FALSE, trim_ws = TRUE)
Parsed with column specification:
cols(
  Code = col_character(),
  Totpor = col_double(),
  PRA = col_double(),
  PV = col_double(),
  Densi = col_double(),
  TenStr = col_double(),
  CO2SBW = col_double(),
  FirTemp = col_integer()
)
>
> ## definisco la funzione che calcola l'intervallo bilaterale con varianza nota
>
> bilat.norm = function(x,sigma,conf) { n = length(x); xbar=mean(x);
+ alpha = 1 - conf;
+ zstar = qnorm(1-alpha/2);
+ SE = sigma/sqrt(n);
+ xbar + c(-zstar*SE,zstar*SE)}
>
> # definisco la funzione che calcola l'intervallo bilaterale con varianza ignota
>
> bilat.stud = function(x,conf) { n = length(x);
+ m = n-1;
+ xbar=mean(x);
+ alpha = 1 - conf;
+ zstar = qt(1-alpha/2, m, lower.tail=TRUE);
```

```
+ SE = sd(x)/sqrt(n);
+ xbar + c(-zstar*SE,zstar*SE)
+ }
>
> # definisco la funzione che calcola l'intervallo bilaterale per la varianza
>
> bilat.chi = function(x,conf) {
+     n = length(x);
+     m = n-1;
+     alpha = 1 - conf;
+     zsup = qchisq(alpha/2, m, lower.tail=TRUE);
+     zinf = qchisq(1 - alpha/2, m, lower.tail=TRUE);
+     SE = sd(x)*sd(x)*m;
+     c(SE/zinf,SE/zsup)
+ }
>
>
> numSummary(table2[,c("Totpor", "PRA", "PV", "Densi", "TenStr", "CO2SBW", "FirTemp")],
+  statistics=c("mean", "sd", "quantiles"), quantiles=c(0,.25,.5,.75,1))
              mean         sd      0%      25%     50%      75%    100%  n NA
Totpor   40.1193548  7.0371760  26.850  36.0550  40.900  44.4200  54.640 31  0
PRA       0.6732581  0.4760389   0.158   0.4220   0.622   0.7305   2.657 31  0
PV       55.3290323 28.5498417  10.200  30.4500  59.400  80.7000  88.600 31  0
Densi     1.6929032  0.1701214   1.340   1.5600   1.680   1.8150   2.020 31  0
TenStr    0.6092258  0.3143682   0.143   0.4065   0.527   0.7165   1.405 31  0
CO2SBW    0.5816667  0.5259152   0.050   0.2900   0.390   0.4950   1.960 30  1
FirTemp 764.8387097 52.9698636 730.000 740.0000 740.000 750.0000 960.000 31  0
>
> bilat.norm(table2$Totpor, 7.04, .9)
[1] 38.03957 42.19914
> bilat.norm(table2$Totpor, 7.04, .95)
[1] 37.64113 42.59758
>
> bilat.stud(table2$Totpor, .9)
[1] 37.97416 42.26455
> bilat.stud(table2$Totpor, .95)
[1] 37.53810 42.70061
>
> bilat.chi(table2$Totpor, .9)
[1] 33.94002 80.33757
> bilat.chi(table2$Totpor, .95)
[1] 31.62366 88.48047
>
```