Appunti del corso di Statistica Matematica

Gadotti Andrea, Nardin Michele, Peruzzetto Marco

Indice

Parte I

Prima parte del corso

Capitolo 1

Introduzione

In questa prima sezione vengono presentati i richiami di teoria della probabilità, affrontati nelle primissime lezioni del corso.

1.1 Funzione generatrice dei momenti

Lezione del 18/02, ultima modifica 09/04, Andrea Gadotti

Definizione 1.1 (fgm)**.** Sia X una variabile casuale, discreta o assolutamente continua. Se esiste $t_0 \in \mathbb{R}_+$ tale per cui $\mathbb{E}(e^{tX}) < +\infty$ per ogni $t \in (-t_0, t_0)$, chiameremo la funzione

$$M_X := \mathbb{E}(e^{tX}) \tag{1.1}$$

funzione generatrice dei momenti di X.

1.1.1 Esempi

Bernoulli Sia $X \sim b(1, p)$, con $p \in (0, 1)$. Si ha:

$$M_X(t) = \mathbb{E}(e^{tX}) = \sum_{x=0}^{1} e^{tx}\mathbb{P}(X = x) = \sum_{x=0}^{1} e^{tx}p^x(1-p)^{1-x} = pe^t + (1-p).$$

Poisson Sia $X \sim \mathcal{P}(\lambda)$, con $\lambda \in \mathbb{R}_+$. Si ha:

$$M_X(t) = \mathbb{E}(e^{tX}) = \sum_{x=0}^{+\infty} e^{tx}\frac{e^{-\lambda}\lambda^x}{x!} = e^{\lambda(e^t-1)}.$$

Gamma Sia $X \sim \mathcal{G}(\alpha, \beta)$: allora, la densità di X è data da

$$f_X(x; \alpha, \beta) := \frac{1}{\Gamma(\alpha)\beta^\alpha}x^{\alpha-1}e^{-\frac{1}{\beta}x}\mathbb{I}_{\mathbb{R}_+}(x), \quad \alpha, \beta \in \mathbb{R}_+$$

dove Γ indica la funzione di Eulero

$$\Gamma(\alpha) := \int_0^{+\infty} x^{\alpha-1}e^{-x}\mathrm{d}x,$$

la quale è tale che $\Gamma(\alpha) = (\alpha - 1)!$ se $\alpha \in \mathbb{N}$.

La generatrice dei momenti di X risulta essere:

$$\begin{aligned} M_X(t) = \mathbb{E}(e^{tX}) &= \int_0^{+\infty} e^{tx} \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-\frac{1}{\beta}x} \mathrm{d}x \\ &= \frac{1}{\Gamma(\alpha)\beta^\alpha} \int_0^{+\infty} e^{-\tau} \frac{\beta^{\alpha-1}}{(1-\beta t)^{\alpha-1}} \tau^{\alpha-1} \frac{\beta}{1-\beta t} \mathrm{d}\tau, \qquad \tau := x\left(\frac{1}{\beta} - t\right) \\ &= \frac{1}{(1-\beta t)^\alpha} \frac{\int_0^{+\infty} \tau^{\alpha-1} e^{-\tau} \mathrm{d}\tau}{\Gamma(\alpha)} = \frac{1}{(1-\beta t)^\alpha}, \qquad t < \frac{1}{\beta}. \end{aligned}$$

1.2 Momenti di una variabile casuale

Definizione 1.2. Se una variabile casuale ammette fgm derivabile infinite volte in un intorno di $t = 0$ e se tutti i suoi momenti sono finiti, allora definiamo il *momento di ordine s non centrato*:

$$\mu'_s := \mathbb{E}(X^s) = \frac{d^s}{dt^s} M_X(t)\bigg|_{t=0}. \tag{1.2}$$

Il *momento di ordine s centrato* in $a \in \mathbb{R}$ è definito come:

$$\mu_s(a) := \mathbb{E}((X-a)^s), \tag{1.3}$$

ovvero $\mu'_s = \mu_s(0)$.

Osservazione 1. É chiaro che $\mu'_1 = \mathbb{E}(X)$. Se non viene specificato altro, chiameremo *momento di ordine s centrato* la quantità

$$\mu_s = \mathbb{E}((X - \mu'_1)^s). \tag{1.4}$$

Inoltre, possiamo osservare che:

$$\mu_2 = \mathbb{E}\left(\left(X - \mu'_1\right)^2\right) = \mathrm{Var}(X) = \mathbb{E}(X^2) - \left(\mathbb{E}\left(X\right)\right)^2 = \mu'_2 - (\mu'_1)^2.$$

Osservazione 2. Grazie a quanto detto in Definizione 1.2 è possibile stabilire la seguente relazione tra momenti centrati e non:

$$\mu_s = \mathbb{E}\left(\left(X - \mu'_1\right)^s\right) = \sum_{m=0}^{s} (-1)^m \binom{s}{m} \mu'_{s-m} (\mu'_1)^m. \tag{1.5}$$

Teorema 1.1. *Date due variabili casuali X e Y indipendenti, aventi densità/massa probabilistica f_X, f_Y e fgm $M_X(t)$, $M_Y(t)$ rispettivamente, vale:*

$$M_{X+Y}(t) = M_X(t) \cdot M_Y(t). \tag{1.6}$$

Teorema 1.2. *Siano X, Y variabili casuali con funzioni di ripartizione $F_X(x)$, $F_Y(y)$ rispettivamente. Siano $M_X(t)$, $M_Y(t)$ fgm di X e Y, esistenti in un intorno aperto $\mathcal{U}$ dell'origine. Sussiste l'equivalenza*

$$M_X(t) = M_Y(t)\ \forall_{t\in\mathcal{U}} \iff F_X(z) = F_Y(z)\ \forall_{z\in\mathbb{R}}. \tag{1.7}$$

Osservazione 3. Il teorema appena visto afferma che, se esiste, la funzione generatrice dei momenti caratterizza la distribuzione della corrispondente variabile casuale.

1.2.1 Esempi

Esempio 1. Siano $(X_1, ..., X_n)$ risultati della replicazione di un esperimento casuale dicotomico ($X_i \sim b(1,p)$). Vogliamo trovare la distribuzione di $S_n := \sum_{i=1}^{n} X_i$. Calcoliamo quindi la sua fgm:

$$M_{S_n}(t) = \mathbb{E}\left(e^{tS_n}\right) = \mathbb{E}\left(e^{t\sum_{i=1}^{n} X_i}\right)$$

$$\overset{TEO2}{=} \prod_{i=1}^{n} \mathbb{E}(e^{tX_i}) = \prod_{i=1}^{n} M_{X_i}(t) = \prod_{i=1}^{n} (pe^t + (1-p)) = (pe^t + (1-p))^n$$

ovvero S_n è distribuita come $b(n,p)$ per il Teorema 2.

Esercizio Ripetere il calcolo precedente supponendo $X_i \sim P(\lambda), \forall i$.

1.3 Famiglia Esponenziale a k parametri

Una famiglia di f densità / f massa è detta essere una Famiglia Esponenziale a k parametri $\theta_1, ..., \theta_k$ se la corrispondente f densità / f massa (che è indicizzata da $\theta_1, ..., \theta_k$) può essere scritta come

$$f_X(x;\theta) = C^*(x)D^*(\theta)\exp\{\sum_{m=1}^{k} A_m(\theta)B_m(x)\}$$

dove $C^*(x)$ è una funzione della sola x, $D^*(\theta)$ è una funzione del solo θ, $A_m(\theta)$ è una funzione del solo θ e $B_m(x)$ è una funzione della sola x.
Esempi

1. $X \sim G(\alpha,\beta) \Longrightarrow f_X(x;\alpha,\beta) = \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-\frac{1}{\beta}x} \mathbb{1}_{\mathbb{R}^+}(x)$, $\alpha > 0$, $\beta > 0$ $\mathbb{1}_{\mathbb{R}^+}$ è detto supporto della distribuzione. Quindi possiamo riscrivere $f_X(x;\alpha,\beta)$ come

   $$f_X(x;\alpha,\beta) = \frac{1}{\Gamma(\alpha)\beta^\alpha} \mathbb{1}_{\mathbb{R}^+}(x) exp((\alpha-1)ln(x) - \frac{1}{\beta}x)$$

   e quindi ponendo $D^*(\alpha,\beta) := \frac{1}{\Gamma(\alpha)\beta^\alpha}$, $C^*(x) := \mathbb{1}_{\mathbb{R}^+}(x)$, $A_1(\alpha,\beta) := (\alpha-1)$, $B_1(x) := ln(x)$, $A_2(\alpha,\beta) := -\frac{1}{\beta}$ e $B_2(x) := x$, otteniamo $G(\alpha,\beta)$ come famiglia esponenziale con $k = 2$.

2. $X \sim b(n,p) \Longrightarrow f_X(x;n,p) = \binom{n}{x} p^x (1-p)^{n-x} \mathbb{1}_{\{0,1,...,n\}}(x)$ con $n \in \mathbb{N}$ noto. Quindi possiamo riscrivere $f_X(x;n,p)$ come

   $$f_X(x;n,p) = \binom{n}{x} \mathbb{1}_{\{0,1,...,n\}}(x)(1-p)^n exp(x\ln(\frac{p}{1-p}))$$

   con $\frac{p}{1-p}$ detto odd ratio o parametro naturale della famiglia esponenziale.
   Quindi ponendo $D^*(p) := (1-p)^n$, $C^*(x) := \binom{n}{x}\mathbb{1}_{\{0,1,...,n\}}(x)$, $A_1(p) := ln(\frac{p}{1-p})$, $B_1(x) := x$, otteniamo $b(n,p)$ come famiglia esponenziale con $k = 1$.

3. X vc con $f_X(x,\vartheta) = \frac{e^{1-x/\vartheta}}{\vartheta}\mathbb{1}_{(\vartheta,\infty)}(x)$: la distribuzione di X non appartiene a famiglia esponenziale. Il fatto che il supporto di f_X dipenda dal parametro ϑ NON permette a f_X di appartenere ud una famiglia esponenziale!

Osservazione 4. Le famiglie di esponenziali hanno interessanti proprietà matematiche (proprietà di regolarità).

Dal punto di vista statistico, ciò si traduce in un'interessante conseguenza: tutta l'informazione contenuta nei dati a disposizione $(X_1, ..., X_n)$ relativa alla funzione $f_X(x;\theta)$ può essere sintetizzata attraverso k quantità (funzioni di $(X_1, ..., X_n)$) che potranno essere impiegate per costruire procedure inferenziali (stima, test per la verifica di ipotesi) riguardanti il parametro θ.
Ovvero, l'appartenenza ad una famiglia esponenziale permette una riduzione dei dati (X_1, ... ,X_n) via B_m.

1.3.1 Trasformazioni di variabili casuali

Lezione del 01/03, ultima modifica 09/04, Michele Nardin

Discrete

Teorema 1. *Sia X una vc con funzione di massa $f_X(x) = P(X = x)$, e sia A_X il suo supporto. Sia W=h(X) una nuova vc. Allora*

$$P(W = w) = \sum_{\{x \in A_X : h(x) = w\}} P(X = x)$$

Esempi

1. Sia $X \sim b(n,p)$ con relativa funzione di massa $f_X(x,p) = \binom{n}{x} p^x (1-p)^{n-x} \mathbb{1}_{0,1,...,n}(x)$, n noto e $p \in (0,1)$.

   Considero quindi $W = n - X$. Come si distribuisce W?

$$P(W = w) = P(X = n - w) = \binom{n}{n-w} p^{n-w}(1-p)^w \mathbb{1}_{0,1,...,n}(w)$$

2. Sia X una vc tale che $f_X(x) = P(X = x) = \left(\frac{1}{2}\right)^x \mathbb{1}_{\mathbb{N}}(x)$, $W = X^3$.

$$P(W = w) = P(X^3 = w) = P(X = \sqrt[3]{w}) = \left(\frac{1}{2}\right)^{\sqrt[3]{w}} \mathbb{1}_{1,8,27,64,...}(w)$$

Assolutamente continue

Teorema 2. *Sia X una variabile casuale (ass continua) con funzione di densità $f_X(x)$ e sia $W = h(X)$, ove h è una funzione monotona. Supponiamo inoltre che $f_X(x)$ sia continua sul supporto di X e che $h^{-1}(w)$ abbia derivata continua sul supporto di W. Allora*

$$f_W(w) = f_X(h^{-1}(w)) \left| \frac{d}{dw} h^{-1}(w) \right| \mathbb{1}_{A_W}(w)$$

Esempio (Standardizzazione di una vc normale) Sia $X \sim N(m, s^2)$. Considero $W = h(X) = \frac{X-m}{s}$. Allora, dato che $h^{-1}(w) = sw + m$, che ha derivata continua su tutto $\mathbb{R}$,

$$f_W(w) = f_X(sw+m)|s| = \frac{e^{\frac{-w^2}{2}}}{\sqrt{2\pi}} = f_{N(0,1)}$$

Teorema 3. *Se $W = h(X)$ ove h è monotona a tratti (un numero di tratti finito k) e valgono le condizioni del teorema precedente (su ogni tratto), allora*

$$f_W(w) = \sum_{n=1}^{k} f_X(h_n^{-1}(w)) \left| \frac{d}{dw} h_n^{-1}(w) \right| \mathbb{1}_{A_W}(w)$$

Esempio (Chi-quadro)
Sia $X \sim N(0,1)$ e $W = h(X) = X^2$. h è monotona sui tratti $A_0 = 0$, $A_1 = (-\infty, 0)$, $A_2 = (0, +\infty)$.
Considero $h_1(x) = x^2$ per $x < 0$ mentre $h_2(x) = x^2$ per $x > 0$.
Trovo che $h_1^{-1}(w) = -\sqrt{w}$ (NB:$h_1^{-1}(w) \in A_1 \forall w \geq 0$), mentre $h_2^{-1}(w) = \sqrt{w}$ (NB:$h_2^{-1}(w) \in A_2 \forall w \geq 0$).
$\frac{d}{dw} h_1^{-1}(w) = -\frac{1}{2\sqrt{w}}$, $\frac{d}{dw} h_2^{-1}(w) = \frac{1}{2\sqrt{w}}$ sono entrambe continue su $\mathbb{R}_+$.

$$f_W(w) = \frac{1}{\sqrt{2\pi}} e^{\frac{-(-\sqrt{w})^2}{2}} \left| \frac{1}{2\sqrt{w}} \right| + \frac{1}{\sqrt{2\pi}} e^{\frac{-(\sqrt{w})^2}{2}} \left| \frac{1}{2\sqrt{w}} \right|$$

$$= \frac{1}{\sqrt{2\pi w}} e^{\frac{-w}{2}} \mathbb{1}_{\mathbb{R}+}(w) = \frac{1}{2^{1/2}\Gamma(1/2)} w^{\frac{1}{2}-1} e^{\frac{-w}{2}}$$

Si riconosce che $W \sim \mathcal{G}(\alpha = 1/2, \beta = 2)$ e si chiama Chi quadrato con $\nu = 1$ gradi di libertà.

In generale, una vc Chi Quadro con $\nu = n$ gradi di libertà è $W = \sum_{i=1}^{n} X_i^2$, ove $X_1, X_2, ..., X_n$ sono vc iid N(0,1). Per il Teorema 2 sulla FGM di una somma di vc iid si trova immediatamente che $W \sim \mathcal{G}(\alpha = n \cdot 1/2, \beta = 2)$.

1.3.2 Convergenze

Convergenza in probabilità

Definizione 1. Sia $\{X_n\}_{n\in\mathbb{N}}$ una successione di variabili casuali e sia X un'altra variabile casuale, tutte definite sullo stesso spazio campionario. Diciamo che X_n converge in probabilità a X (scriviamo $X_n \xrightarrow{p} X$) se $\forall \varepsilon > 0$

$$\lim_{n\to\infty} P(|X_n - X| \geq \varepsilon) = 0$$

Osservazione 5. Se $X_n \xrightarrow{p} X$ diciamo che la "massa" della differenza $|X_n - X|$ converge a 0. Inoltre, quando scriviamo $X_n \xrightarrow{p} X$, stiamo sottintendendo tutta la parte iniziale della definizione precendete, cioè il "sia $\{X_n\}_{n\in\mathbb{N}}$ una successione di variabili casuali...".

Teorema 4. *Alcuni risultati utili:*

1. *Supponiamo che $X_n \xrightarrow{p} X$ e $Y_n \xrightarrow{p} Y$. Allora $X_n + Y_n \xrightarrow{p} X + Y$*

2. *Supponiamo che* $X_n \xrightarrow{p} X$ *e sia a una costante. Allora* $aX_n \xrightarrow{p} aX$

3. *Supponiamo che* $X_n \xrightarrow{p} a$ *costante, e sia g una funzione reale continua in a. Allora* $g(X_n) \xrightarrow{p} g(a)$

4. *(Corollario di 3.) Se* $X_n \xrightarrow{p} a$*, allora* $X_n^2 \xrightarrow{p} a^2$*,* $\frac{1}{X_n} \xrightarrow{p} \frac{1}{a}$ *(se* $a \neq 0$*),* $\sqrt{X_n} \xrightarrow{p} \sqrt{a}$ *(*$a \geq 0$*).*

5. $X_n \xrightarrow{p} X$ *e* $Y_n \xrightarrow{p} Y$ *allora* $X_n Y_n \xrightarrow{p} XY$

Convergenza in distribuzione

Definizione 2. Sia $\{X_n\}_{n\in\mathbb{N}}$ una successione di variabili casuali e sia X un'altra variabile casuale, tutte definite sullo stesso spazio campionario.

Siano F_{X_n} e F_X le relative funzioni di ripartizione (dette anche "di distribuzione"). Sia $C(F_X)$ l'insieme dei punti ove F_X è continua. Diciamo che X_n converge in distribuzione (o in legge) a X (scriviamo $X_n \xrightarrow{d} X$) se

$$\lim_{n\to\infty} F_{X_n}(x) = F_X(x) \quad \forall x \in C(F_X)$$

Teorema 5. *Se* $X_n \xrightarrow{p} X$ *allora* $X_n \xrightarrow{d} X$*.*

Osservazione 6. Il contrario in generale non vale, tranne nel caso in cui X è una vc degenere (cioè costante).

Teorema 6. *Supponiamo che* $X_n \xrightarrow{d} X$ *e sia g una funzione continua sul supporto di X. Allora* $g(X_n) \xrightarrow{d} g(X)$

Teorema 7 (Slutsky)**.** *Supponiamo che* $X_n \xrightarrow{d} X$*,* $A_n \xrightarrow{p} a$ *costante e* $B_n \xrightarrow{p} b$ *costante. Allora* $A_n + B_n X_n \xrightarrow{d} a + bX$

1.3.3 Teoria asintotica

Lezione del 04/03, ultima modifica 09/04, Michele Nardin

Teorema 8. *(*Δ*-method) Sia* $\{X_n\}_{n\in\ \mathbb{N}}$ *una successione di vc tale che* $\sqrt{n}(X_n - \vartheta) \xrightarrow{d} N(0, \sigma^2)$*. Supponiamo che una funzione g(X) sia derivabile in* ϑ *e che* $g'(\vartheta) \neq 0$*. Allora*

$$\sqrt{n}(g(X_n) - g(\vartheta)) \xrightarrow{d} N(0, \sigma^2(g'(\vartheta))^2)$$

Esempio Considero

$$Y_n = \frac{\chi_n^2 - n}{\sqrt{2n}} = \sqrt{n}\left(\frac{\chi_n^2}{\sqrt{2n}} - \frac{1}{\sqrt{2}}\right)$$

ove χ_n^2 è la chiquadro con n gradi di libertà. Ricordiamo che $\mathbb{E}(\chi_n^2) = n$ e che $Var(\chi_n^2) = 2n$ (discende dal fatto che $\chi_n^2 \sim \mathcal{G}(\alpha = n/2, \beta = 2)$. Affermiamo che $Y_n \xrightarrow{d} N(0,1)$. Infatti:

$$Y_n = \frac{\chi_n^2 - n}{\sqrt{2n}} = \frac{\sum_{i=1}^n X_i^2 - n \cdot 1}{\sqrt{n}\sqrt{2}}$$

dove $X_i \sim N(0,1)$, e quindi $X_i^2 \sim \chi_1^2$, quindi le X_i^2 hanno media $\mu = 1$ e varianza $\sigma^2 = 2$. Quindi per il Teorema centrale del Limite (vedi sotto) si ha quanto voluto.
Scrivendo ora Y_n nella forma $Y_n = \sqrt{n}\left(\frac{\chi_n^2}{\sqrt{2}n} - \frac{1}{\sqrt{2}}\right)$ riconosciamo che la prima parte delle ipotesi del Δ-method sono soddisfatte. Considero quindi $g(t) = \sqrt{t}$, che è derivabile in $\vartheta = 1/\sqrt{2}$, $g'(t) = \frac{1}{2\sqrt{t}}|_{\vartheta=1/\sqrt{2}} = 2^{-3/4}$. Allora

$$\sqrt{n}(g\left(\frac{\chi_n^2}{\sqrt{2}n}\right) - g(\vartheta)) = \sqrt{n}\left(\sqrt{\frac{\chi_n^2}{\sqrt{2}n}} - \sqrt{\frac{1}{\sqrt{2}}}\right) \xrightarrow{d} N(0, 1^2 \cdot 2^{-3/2})$$

Teorema 9. *(Teorema centrale del limite) Siano $X_1, ...X_n$ vc iid dotate di media μ e varianza finita σ^2. Allora*

$$\frac{\sum_{i=1}^n X_i - n\mu}{\sqrt{n}\cdot\sigma} = \frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma} \xrightarrow{d} N(0,1)$$

con $\overline{X}_n$ media aritmetica delle X_i.

Esempi/Applicazioni

1. $X_n \sim b(n,p)$, $X_n \overset{a}{\sim} N(np, np(1-p))$ (ricordiamo che $X_n \sim \sum_{i=1}^n b_i$, ove $b_i \sim b(1,p)$). Quando scriviamo $\overset{a}{\sim}$ stiamo considerando un "andamento asintotico", ossia sottintendiamo un'approssimazione (via via migliore con l'aumentare di n) giustificata dal TLC (il senso è che per n 'grandi' la distribuzione 'funziona circa così').

2. $X_1, ..., X_n$ vc $P(\lambda = 1)$. Considero $Y_n = \sum X_i$. Dato che $Y_n \overset{a}{\sim} N(n\lambda, n\lambda)$ e $\lambda = 1$, $\bar{Y}_n := \frac{Y_n}{n} \overset{a}{\sim} N(1, 1/n)$

   Considero quindi $W_n = \sqrt{n}(\frac{Y_n}{n} - 1) = \frac{\frac{Y_n}{n} - 1}{1/\sqrt{n}} = \frac{\bar{Y}_n - \mathbb{E}(\bar{Y}_n)}{\sqrt{Var(\bar{Y}_n)}}$, trovo che $W_n \overset{a}{\sim} N(0,1)$

Teorema 10. *Sia $\{X_n\}$ una succ di vc iid, ognuna con con FGM $M_{X_n}(t)$ definita e $< \infty$ per $t \in (-h, h)$, e sia X un'altra vc con FGM $M_X(t)$ definita e $< \infty$ per $t \in (-h_1, h_1)$, $h_1 \leq h$. Se*

$$\lim_{n\to+\infty} M_{X_n}(t) = M_X(t) \quad \forall |t| \leq h_1$$

allora $X_n \xrightarrow{d} X$.

Applicazione

1. Sia $X_n \sim b(n,p)$. Ricordiamo che $X_n = \sum X_i$ ove $X_i \sim b(1,p)$, ed inoltre $\mu = \mathbb{E}(X) = np$. Siccome $M_{X_n}(t) = \mathbb{E}(e^{tX_n}) = [(1-p) + pe^t]^n = [1 + \frac{\mu}{n}(e^t - 1)]^n$,

   $$M_{X_n}(t) \overset{n\to\infty}{\longrightarrow} e^{\mu(e^t-1)}$$

   che è la FGM di una Poisson di parametro μ, ovvero $X_n \xrightarrow{d} \mathcal{P}(n,p)$.

1.4 Approccio applicativo alla Statistica Matematica

Questa sezione corrisponde alla parte di corso svolta dalla seconda settimana di marzo fino a metà aprile, che riguarda gli aspetti pratici della statistica: verranno introdotte le statistiche d'ordine, gli intervalli di confidenza e i test per verifiche d'ipotesi.

Definizione 3. (Campione Casuale) Il vettore casuale $(X_1, ..., X_n)$ si dice Campione Casuale relativamente ad una vc $X \sim F_X(x, \vartheta)$ se i suoi elementi sono vc i.i.d.

Osservazione Il fatto che le vc siano i.i.d. implica che

$$F_{X_1,...,X_n}(X_1, ..., X_n) = \prod_{i=1}^{n} F_{X_i}(X_i)$$

e

$$f_{X_1,...,X_n}(X_1, ..., X_n) = \prod_{i=1}^{n} f_{X_i}(X_i)$$

Definizione 4. (Statistica) Sia $(X_1, ..., X_n)$ un campione casuale da una distribuzione associata alla vc X, e sia Ω lo spazio campionario di $(X_1, ..., X_n)$
Ogni funzione

$$T(X_1, ..., X_n) : \Omega \longrightarrow \mathbb{R}^k$$

che NON dipende da parametri incogniti è detta Statistica.

1.4.1 Statistiche d'ordine

Lezioni 08 e 11 Marzo, ultima modifica 21/03, Scritte da: Marco Peruzzetto

Definizione 5. Sia $(X_1, \ldots, X_n)$ un campione casuale con distribuzione $F_X(x, \theta)$, densità $f_X(x, \theta)$ e supporto $\text{supp}\{X\} \coloneqq (a, b) \subset \mathbb{R}$ ove $X \in \{X_1, \ldots, X_n\}$ e $-\infty \leq a < b \leq +\infty$. Definiamo ricorsivamente le seguenti variabili casuali:

- $X_{(1)} \coloneqq \min(\{X_1, \ldots, X_n\})$;
- $X_{(i)} \coloneqq \min\left(\{X_1, \ldots, X_n\} \setminus \{X_{(1)}, \ldots, X_{(i-1)}\}\right)$ $\forall 1 < i \leq n$.

Chiameremo allora $X_{(i)}$ la i-esima *Statistica d'Ordine* del campione.

Osservazione: La statistica d'ordine consiste semplicemente nel vettore per il quale le variabili casuali vengono appunto ordinate, in base al valore che assumono in un determinato punto del loro dominio comune, in ordine crescente. In particolare $X_{(i)}$ sarà l' i-esima variabile più piccola. Naturalmente, se il campione ha lunghezza n, allora $X_{(n)} = \max(\{X_1, \ldots, X_n\})$. Osserviamo che la funzione $(X_1, \ldots, X_n) \longmapsto \left(X_{(1)}, \ldots, X_{(n)}\right)$ è essa stessa una Statistica.

Teorema 11. *Sia $(X_1, \ldots, X_n)$ un campione casuale come sopra. Allora si ottiene $\forall$ $1 \leq m \leq n$ che la densità dell' m-esima statistica d'ordine è data da:*

$$f_{X_{(m)}}(x, \theta) = \frac{n!}{(m-1)!(n-m)!} f_X(x, \theta) \cdot F_X(x, \theta)^{m-1} \cdot \left(1 - F_X(x, \theta)\right)^{n-m}$$

Daremo due dimostrazioni, la seconda più bella della prima.

Dimostrazione. (a cura di Marco Perruzzetto) Innanzi tutto si ha che il supporto (a, b) può essere partizionato in n parti, per cui evidentemente si ha:

$$f_{(X_{(1)},\dots,X_{(n)})}(x_{(1)},\dots,x_{(n)},\theta) = \begin{cases} n!\prod_{i=1}^{n} f_X(x_{(i)},\theta) & \text{se } a < x_{(1)} < x_{(2)} < \dots < x_{(n)} < b \\ 0 & \text{altrimenti.} \end{cases}$$

ove la produttoria è giustificata dal fatto che le variabili sono tutte indipendenti e che devono essere ciascuna minore dell'altra per l'ordinamento assegnato; il coefficiente fattoriale è presente poiché le n parti dell'intervallo (a, b) possono essere assegnate alle n variabili in tale numero di modi, dato che ciascuna X_i $\forall 1 \leq i \leq n$ ha la stessa distribuzione.
Adesso per trovare la distribuzione di ciascuna $X_{(m)}$ sarà dunque sufficiente integrare $f_{(X_{(1)},\dots,X_{(n)})}$ nei domini possibili di tutte le altre funzioni di distribuzione di ciascuna $X_{(i)}$ con $i \neq m$. In particolare, ciascuna $f_{X_{(i)}}$ per $i < m$ dovrà assumere a piacere valori necessariamente inferiori a $f_{X_{(m)}}$, viceversa ogni $f_{X_{(i)}}$ per $i > m$ dovrà assumere valori obbligatoriamente superiori a quelli di $f_{X_{(m)}}$ in ogni punto. Ricordando allora che possiamo scrivere la distribuzione come $\int_a^x f_X(\theta, t)dt = F_X(\theta, x)$ essendo la densità la derivata della funzione di distribuzione, otterremo quindi che $\forall a < x_{(m)} < b$ la distribuzione sarà data da:

$$\begin{aligned} f_{X(m)}(x_{(m)},\theta) = & \\ = & \int_a^{x_{(2)}} dx_{(1)} \cdots \int_a^{x_{(m)}} dx_{(m-1)} \int_{x_{(m)}}^{b} dx_{(m+1)} \cdots \int_{x_{(n-1)}}^{b} dx_{(n)} f_{(X_{(1)},\dots,X_{(n)})}(x_{(1)},\dots,x_{(n)},\theta) \\ = & \int_a^{x_{(2)}} dx_{(1)} \cdots \int_a^{x_{(m)}} dx_{(m-1)} \int_{x_{(m)}}^{b} dx_{(m+1)} \cdots \int_{x_{(n-1)}}^{b} dx_{(n)} n! \prod_{i=1}^{n} f_X(x_{(i)},\theta) \\ = & f_X(x_{(m)}) \int_a^{x_{(2)}} dx_{(1)} \cdots \int_a^{x_{(m)}} dx_{(m-1)} \int_{x_{(m)}}^{b} dx_{(m+1)} \cdots \int_{x_{(n-1)}}^{b} dx_{(n)} n! \prod_{i=1, i\neq m}^{n} f_X(x_{(i)},\theta) \\ = & \frac{n!}{(m-1)!(n-m)!} f_X(x,\theta) \cdot F_X(x,\theta)^{m-1} \cdot \big(1 - F_X(x,\theta)\big)^{n-m}, \end{aligned}$$

dove è stato usato il fatto che $\int_a^b F_X^{\alpha}(\theta, t) f_X(\theta, t)dt = \frac{F_X^{\alpha+1}}{\alpha+1}$, $\forall \alpha \neq -1$. □

Dimostrazione. Sia Ω il dominio comune del campione casuale. Definiamo per $x \in \mathbb{R}$ la nuova variabile casuale Y_x come:

$$\begin{aligned} \Omega &\longrightarrow \{0,\dots,n\} \\ Y_x(\omega) \coloneqq \sum_{i=1}^{n} \mathbb{1}_{\{X_i(\omega)\leq x\}}(\omega) &= \#\big\{i \in \{1,\dots,n\} : X_i \leq x\big\}, \end{aligned}$$

funzione che, per così dire, "conta" il numero di variabili casuali X_i che non superano x. Si vede immediatamente che $\forall 1 \leq m \leq n$, si ha la distribuzione

$$\begin{aligned} F_{X_{(m)}}(\theta, x) &= \mathbb{P}[X_{(m)} \leq x] = \mathbb{P}[\text{almeno } X_{(1)}, ..., X_{(m)} \text{ stanno sotto } x] = \\ &= \mathbb{P}[Y_x \geq m] = \sum_{k=m}^{n} \mathbb{P}[Y_x = k] = \\ &= \sum_{k=m}^{n} \binom{n}{k} F_X^k(\theta, x)\big(1 - F_X(\theta, x)\big)^{n-k}. \end{aligned}$$

Come nella prima dimostrazione usiamo il fatto che la densità si può vedere come derivata della funzione di ripartizione. Ne segue che per calcolare la densità sarà sufficiente calcolare la derivata in ciascun punto x della distribuzione appena trovata. In particolare si potrà vedere che coesisteranno il termine che vogliamo ottenere con altre due sommatorie, che tuttavia si elidono l'una con l'altra lasciando quindi la relazione espressa dal teorema. Si ha infatti che:

$$
\begin{aligned}
f_{(m)}(\theta,x) &= \frac{\partial}{\partial x}F_{X_{(m)}}(\theta,x) = \\
&= \sum_{k=m}^{n}\binom{n}{k}\cdot f_X(\theta,x)\Big\{kF_X^{k-1}(\theta,x)\big(1-F_X(\theta,x)\big)^{n-k}-(n-k)F_X^{k}(\theta,x)\big(1-F_X(\theta,x)\big)^{n-k-1}\Big\} \\
&= m\binom{n}{m}\cdot f_X(\theta,x)F_X^{m-1}(\theta,x)\big(1-F_X(\theta,x)\big)^{n-m}+ \\
&\quad \sum_{k=m+1}^{n}k\binom{n}{k}f_X(\theta,x)F_X^{k-1}(\theta,x)\big(1-F_X(\theta,x)\big)^{n-k}- \\
&\quad \sum_{k=m}^{n-1}(n-k)\binom{n}{k}f_X(\theta,x)F_X^{k}(\theta,x)\big(1-F_X(\theta,x)\big)^{n-k-1} \\
&= \frac{n!}{(m-1)!(n-k)!}\cdot f_X(\theta,x)F_X^{m-1}(\theta,x)\big(1-F_X(\theta,x)\big)^{n-m}+ \\
&\quad \sum_{j=m}^{n-1}(j+1)\binom{n}{j+1}f_X(\theta,x)F_X^{j}(\theta,x)\big(1-F_X(\theta,x)\big)^{n-j-1}- \\
&\quad \sum_{k=m}^{n-1}(n-k)\binom{n}{k}f_X(\theta,x)F_X^{k}(\theta,x)\big(1-F_X(\theta,x)\big)^{n-k-1} \\
&= \frac{n!}{(m-1)!(n-k)!}\cdot f_X(\theta,x)F_X^{m-1}(\theta,x)\big(1-F_X(\theta,x)\big)^{n-m}+ \\
&\quad \sum_{j=m}^{n-1}\frac{n!}{j!(n-j-1)!}f_X(\theta,x)F_X^{j}(\theta,x)\big(1-F_X(\theta,x)\big)^{n-j-1}- \\
&\quad \sum_{k=m}^{n-1}\frac{n!}{k!(n-k-1)!}f_X(\theta,x)F_X^{k}(\theta,x)\big(1-F_X(\theta,x)\big)^{n-k-1} \\
&= \frac{n!}{(m-1)!(n-k)!}\cdot f_X(\theta,x)F_X^{m-1}(\theta,x)\big(1-F_X(\theta,x)\big)^{n-m}.
\end{aligned}
$$

□

Definizione. Sia $\big(X_{(1)},\dots,X_{(n)}\big)$ una statistica d'ordine di un campione casuale. Allora possiamo definire le nuove seguenti variabili:

- $X_{(n)}-X_{(1)}$, detta *Range* oppure *Misura di Dispersione*;
- $\frac{X_{(1)}+X_{(n)}}{2}$ detta *Mid Range* oppure *Misura di Centralità*;
- $$\left.\begin{array}{rl} \forall n \text{ pari} & \frac{X_{(\frac{n}{2})}+X_{(\frac{n}{2}+1)}}{2} \\ \forall n \text{ dispari} & X_{(\frac{n+1}{2})} \end{array}\right\} \text{dette ciascuna } \textit{Mediana campionaria};$$
- Sia $\frac{1}{2(n+1)} < p < 1-\frac{1}{2(n+1)}$, che possiamo in ogni caso pensare come $0<p<1$ per n molto grande. A questo punto possiamo definire l'intero $k_p \coloneqq \big\lfloor p(n+1)\big\rfloor + \big\lfloor 2\big(p(n+$

$1) - 2\lfloor p(n+1)\rfloor\big)\big\rfloor$, che risulta essere così ben definito in quanto compreso tra 1 e n e restituisce l'approssimazione all'intero più vicino al variare di p del reale $p(n+1)$.

- A questo punto, se scegliamo $\xi_p \in F_X^{-1}(p)$, chiameremo ξ_p *Quantile di popolazione* di ordine p. In seguito troveremo utile stimare tale valore. Perciò introduciamo la variabile casuale ad esso collegata $X_{(k_p)}$, detta *Quantile campionario* di ordine p. Se $p = \frac{i}{m}$, allora $X_{(k_p)}$ è detta anche i-esimo m-ile campionario. In particolare con Q_1 e Q_3 si indicano rispettivamente il primo e il terzo quartile.
  Intuitivamente, $X_{(k_p)}$ mi dà la v.c. che sta al k_p-esimo posto, ovvero al $p(n+1)$-esimo posto (se $p(n+1)$ è intero). Ad esempio, se $p = \frac{1}{3}$, $X_{(k_{\frac{1}{3}})}$ è la v.c. che nel vettore ordinato sta alla posizione $\frac{n+1}{3}$.
- Le variabili $LF := Q_1 - h$ e $UF := h + Q_3$, ove $h := \frac{3}{2}(Q_3 - Q_1)$ sono dette rispettivamente *Lower* e *Upper Fence*.

Osservazione: Osserviamo che più la misura di centralità si discosta dalla mediana, più vi è asimmetria nella funzione di densità f (*i.e.:* una funzione di distribuzione è simmetrica $:\iff \exists x_0 \in \mathbb{R} : f(x_0 + x) = f(x_0 - x), \forall x \in \mathrm{Dom}(f)$.) Inoltre, ponendo che la funzione di ripartizione sia iniettiva e la funzione di densità sia simmetrica, si vede immediatamente che la media di popolazione, ovvero il quantile di popolazione di ordine $p = \frac{1}{2}$ coincide con il valore di aspettazione della variabile casuale, il quale a sua volta deve coincidere con x_0.

Dato un campione casuale di parametro $\theta \in \mathbb{R}$ fissato, sappiamo che una qualsiasi funzione di statistiche su tali variabili è, proprio per definizione, uno stimatore del parametro θ. L'esistenza di un'infinità non numerabile di stimatori è sicuramente un problema da ovviare in merito alla scelta tra essi di uno stimatore che effettivamente permetta di stimare il più correttamente possibile il parametro θ. Cercheremo dunque di individuare alcune proprietà che possano effettivamente giustificare la scelta di un determinato stimatore, affinché esso risulti il più possibile affidabile.

Definizione. Sia $(X_1 \dots, X_n)$ un campione di parametro θ e $T_n(X_1, \dots, X_n)$ uno stimatore. La funzione $B_\theta[T_n(X_1, \dots, X_n)] := \mathbb{E}_\theta[T_n(X_1, \dots, X_n)] - \theta$ si dice *distorsione* di T_n (nota: con $\mathbb{E}_\theta$ formalmente intendiamo semplicemente $\mathbb{E}$). In particolare T_n si dirà *non distorto* se e solo se la sua distorsione è nulla $\forall \theta \in \mathbb{R}$ (nel senso che uno stimatore -e.g. la media campionaria- non può stimare bene solo "alcune" medie, ma qualsiasi media reale, ad esempio la media di qualsiasi normale centrata in qualsiasi punto). Altrimenti si dice *distorto*. Se infine si ottiene che $\lim_{n\to+\infty} B_\theta[T_n(X_1, \dots, X_n)] = 0$, T_n si dice *asintoticamente non distorto*.

Esempio Sia $(X_{(1)}, \dots, X_{(n)})$ un campione casuale con distribuzione simmetrica (senza perdita di generalità, la assumiamo simmetrica rispetto all'origine) e scegliamo come stimatore T_n proprio la mediana campionaria. È chiaro innanzi tutto che essa in generale gode delle seguenti due proprietà:

- $\forall b \in \mathbb{R}$, $T_n(X_1 + b, \dots, X_n + b) = T_n(X_1, \dots, X_n) + b$;
- $T_n(-X_1, \dots, -X_n) = -T_n(X_1, \dots, X_n)$.

Abbiamo inoltre che la distribuzione di $(X_1, \ldots, X_n)$ e del vettore $(-X_1, \ldots, -X_n)$ coincidono (ricordando che l'origine è il centro di simmetria). Si avrà dunque:

$$\begin{aligned}\mathbb{E}[T_n] &= \mathbb{E}[T_n(X_1, \ldots, X_n)] = \mathbb{E}[T_n(-X_1 \ldots, -X_n)] \\ &= \mathbb{E}[-T_n(X_1, \ldots, X_n)] = -\mathbb{E}[T_n(X_1 \ldots, X_n)] \\ &= -\mathbb{E}[T_n]\end{aligned}$$

perciò, in definitiva, $2\mathbb{E}[T_n] = 0$ ovvero $\mathbb{E}[T_n] = 0$. Quindi, nel caso di una distribuzione simmetrica, la media campionaria è uno stimatore non distorto del valore di aspettazione, del punto di simmetria e della media della popolazione (dato che tutti loro nel nostro caso coincidono).

Esempio Sia $(X_1, \ldots, X_n)$ un campione casuale di parametro $\theta \in \mathbb{R}$ fissato e con $\mu \coloneqq \mathbb{E}[X]$, $\sigma^2 \coloneqq \mathrm{Var}[X]$. Vogliamo provare a calcolare la distorsione di due stimatori "classici":

1. Scegliamo come stimatore la *media campionaria* $\overline{X}_n \coloneqq \frac{1}{n}\sum_{i=1}^n X_i$. Allora $B_\theta[\overline{X}_n] = \mathbb{E}_\theta[\overline{X}_n] - \theta = \frac{1}{n}\sum_{i=1}^n \mathbb{E}_\theta[X_i] - \theta = \frac{1}{n}\cdot n\mathbb{E}_\theta[X] = \mu - \theta$. Dunque la distorsione è costante $\forall n \in \mathbb{N}$. In particolare è uno stimatore non distorto per il valore di aspettazione μ. Possiamo calcolare facilmente anche la varianza di $\overline{X}_n$ che risulta essere $\frac{\sigma^2}{n}$. La media campionaria si rivela essere quindi un buon stimatore. (nota: nel calcolo della varianza stiamo trattando lo stimatore come una variabile casuale essa stessa, quindi più la varianza è piccola migliore è lo stimatore).

2. Prendiamo ora come stimatore la *varianza campionaria*, data dalla variabile $S_n^2 \coloneqq \frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X}_n)^2$. Allora:
   $\mathbb{E}[S_n^2] = \frac{1}{n-1}\sum_{i=1}^n \mathbb{E}[(X_i - \overline{X}_n)^2] = \frac{1}{n-1}\sum_{i=1}^n \mathbb{E}\big[\,\big((X_i - \mu) - (\overline{X}_n - \mu)\big)^2\,\big] =$
   $= \frac{1}{n-1}\left(\sum_{i=1}^n \mathbb{E}[(X_i - \mu)^2] - \sum_{i=1}^n \mathbb{E}[(\overline{X}_n - \mu)^2]\right) = \frac{1}{n-1}\cdot(n-1)\sigma^2 = \sigma^2$. Perciò S_n^2 è uno stimatore non distorto di σ^2. Notiamo che lo stimatore $S_n^* \coloneqq \frac{1}{n}\sum_{i=1}^n (X_i - \overline{X}_n)^2$ avrebbe distorsione $-\frac{\sigma^2}{n}$, e dunque è peggiore della varianza campionaria, anche se è asintoticamente non distorto.

Calcoleremo adesso la varianza della varianza campionaria. Assumiamo per il momento che il campione provenga da una distribuzione normale $N(\mu, \sigma^2)$. In tal caso mostriamo che $\frac{n-1}{\sigma^2}S_n^2 \sim \chi^2_{n-1}$ e dunque si avrà subito $\mathrm{Var}[S_n^2] = \frac{2\sigma^4}{n-1}$. Infatti, $\frac{n-1}{\sigma^2}S_n^2 = \sum_{i=1}^n \left(\frac{X_i - \overline{X}_n}{\sigma}\right)^2 =$ $\sum_{i=1}^n \left(\frac{X_i-\mu}{\sigma} - \frac{\overline{X}_n - \mu}{\sigma}\right)^2 = \sum_{i=1}^n \left(\frac{X_i-\mu}{\sigma}\right)^2 - n\left(\frac{\overline{X}_n-\mu}{\sigma}\right)^2 = \sum_{i=1}^n \left(\frac{X_i-\mu}{\sigma}\right)^2 - \left(\frac{\overline{X}_n-\mu}{\frac{\sigma}{\sqrt{n}}}\right)^2 \Rightarrow$ $\sum_{i=1}^n \big(\sim \chi^2_1\big) - \big(\sim \chi^2_1\big) \Rightarrow \frac{n-1}{\sigma^2}S_n^2 \sim \chi^2_{n-1}$, ove abbiamo usato il seguente teorema:

Teorema 12. *Sia $(X_1, \ldots, X_n)$ un campione casuale ove la funzione generatrice di ciascuna X_i, $1 \le i \le n$ è $M_X(t)$. Allora $M_{\overline{X}_n}(t) = (M_X(\frac{t}{n}))^n$.*

per mostrare che $\overline{X}_n \sim N(\mu, \frac{\sigma^2}{n})$. Infatti:

$M_{\overline{X}_n}(t) = (M_X(\frac{t}{n}))^n = \left(e^{\mu\frac{t}{n} + \frac{\sigma^2 t^2}{2n^2}}\right)^n = e^{\mu t + \frac{\sigma^2}{n}\cdot\frac{t}{2}}$, da cui la tesi.

Teorema 13. *Sia $(X_1, \dots, X_n)$ un campione casuale da una popolazione con distribuzione discreta o assolutamente continua dove la densità associata sia della forma $f(x, \theta) = C(x)D(\theta)\exp\{\sum_{m=1}^{k} A_m(\theta)B_m(x)\}$ con k naturale positivo. Siano $T_1, \dots, T_k$ statistiche definite $\forall 1 \leq m \leq k$ da $T_m(X_1, \dots, X_n) \coloneqq \sum_{i=1}^{n} B_m(X_i)$. Allora la distribuzione di $(T_1, \dots, T_k)$ sarà ancora della forma esponenziale:*

$$f_{(T_1,\dots,T_k)}(\theta, t_1, \dots, t_k) = C(t_1, \dots, t_k)D(\theta)^n exp\left\{ \sum_{i=1}^{k} A_m(\theta)t_m \right\}$$

.

Esempio. Sia $(X_1, \dots, X_n)$ un campione casuale. Supponiamo che $X \sim \text{Bin}(1, p)$. Allora la densità sarà discreta, ossia sarà $f(p, x) = \mathbb{P}[X = x] = \mathbb{1}_{\{0,1\}}(x)(1-p)\exp\left\{x \log\left(\frac{p}{1-p}\right)\right\}$. Applicando il teorema otteniamo $T_1(X_1, \dots, X_n) = \sum_{i=1}^{n} B_1(X_i) = \sum_{i=1}^{n} X_i$ da cui si deduce subito che $T_1 \sim \text{Bin}(n, p)$. In particolare possiamo scriverne la densità: $f_{T_1}(p, t_1) = \mathbb{1}_{\{0,\dots,n\}}(t_1)\binom{n}{t_1}(1-p)^n \exp\{t_1 \log\left(\frac{p}{1-p}\right)\}$.
Esempio. Sia $(X_1, \dots, X_n)$ un campione casuale e supponiamo che il nostro campione casuale abbia distribuzione uniforme $\text{Unif}([0, \theta])$. Vogliamo stimare θ. Supponiamo che, essendo θ il massimo valore che ciascuna variabile può assumere, un plausibile buon stimatore possa essere proprio il massimo della statistica ordinata, ovvero $T_n(X_1, \dots, X_n) \coloneqq X_{(n)}$. Per il teorema di pagina 8, ne conosciamo già la distribuzione:

$$f_{T_n}(\theta, x) = \frac{n!}{(n-1)!(n-n)!}\frac{1}{\theta}\left(\frac{x}{\theta}\right)^{n-1}\left(1 - \frac{x}{\theta}\right)^{n-n} = \frac{n}{\theta^n}x^{n-1}$$

Allora $\mathbb{E}[T_n] = \int_0^\theta \frac{n}{\theta^n}x^n dx = \frac{n}{n+1}\theta \neq \theta \implies B_\theta[T_n] = \frac{-\theta}{n+1}$. Ne segue che è distorto, ma asintoticamente non distorto per θ. Possiamo anche calcolarne la varianza: $\text{Var}[T_n] = \mathbb{E}[T_n^2] - \mathbb{E}[T_n]^2 = \int_0^\theta \frac{n}{\theta^n}x^{n+1}dx - \frac{n}{n+1} = \frac{n\theta^2}{(n+1)^2(n+2)} \xrightarrow[n\to\infty]{} 0$. Perciò il massimo $X_{(n)}$ rimane in ogni caso uno stimatore affidabile. Osserviamo che possiamo tuttavia introdurre un nuovo stimatore che ci assicura la non distorsione, ovvero $T_n^* \coloneqq \frac{n+1}{n}T$, che possiede le proprietà cercate.

Definizione. Sia $(X_1, \dots, X_n)$ un campione casuale con distribuzione $F(\theta, x)$, ove $\theta \in \Theta \subset \mathbb{R}$. Sia poi T_n una statistica $\forall n \in \mathbb{N}$. Diremo T_n essere uno stimatore *consistente* di $\theta :\iff T_n(X_1, \dots, X_n) \xrightarrow[n\to\infty]{\mathbb{P}} \theta$.

Esempio. Sia $(X_1, \dots, X_n)$ un campione casuale, ove $X \in \mathcal{L}^2(\mathbb{R})$. Indichiamo come al solito media e varianza rispettivamente con μ e σ^2. Allora abbiamo:

1. La media campionaria $\overline{X}_n \coloneqq \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow[n\to\infty]{\mathbb{P}} \mu$, grazie alla legge debole dei grandi numeri poiché $\lim_{n\to+\infty} \mathbb{P}[(\overline{X}_n - \mu) > \varepsilon] = 0$, $\forall \varepsilon > 0$.
2. Consideriamo adesso la varianza campionaria

$$S_n^2 \coloneqq \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2 = \frac{n}{n-1}\left(\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \overline{X}_n^2\right)$$

. Abbiamo ora i seguenti tre termini:

- $\lim_{n\to+\infty} \frac{n}{n-1} = 1$, un semplice limite;
- $\frac{1}{n}\sum_{i=1}^{n} X_i^2 \xrightarrow[n\to\infty]{\mathbb{P}} \mathbb{E}[X^2]$, ancora grazie alla legge debole dei grandi numeri e al fatto che X^2 rimane ancora sommabile;
- $\overline{X}_n^2 \xrightarrow[n\to\infty]{\mathbb{P}} \mu^2 = \mathbb{E}[X]^2$ grazie al Teorema 4 sulla convergenza.

Ne segue quindi che $S_n^2 \xrightarrow[n\to\infty]{\mathbb{P}} \sigma^2$, sempre per i teoremi sulla convergenza di somma, prodotto e prodotto per costanti di variabili casuali.

3. Consideriamo ancora il campione casuale distribuito uniformemente $\mathrm{Unif}([0,\theta])$ con stimatore $T_n(X_1\ldots,X_n) \coloneqq X_{(n)}$. Troviamo che anch'esso è consistente per la stima del massimo. Infatti, $\mathbb{P}[|T_n-\theta|>\varepsilon] = \mathbb{P}[\theta-T_n>\varepsilon] = \mathbb{P}[X_{(n)} \le \theta-\varepsilon] = F_{X_{(n)}}(\theta-\varepsilon) = \left(1-\frac{\varepsilon}{\theta}\right)^n \xrightarrow[n\to\infty]{} 0$. Allo stesso modo si può verificare che anche T_n^* è consistente per θ.

Definizione. Sia $(X_1,\ldots,X_n)$ un campione casuale e $T_n : \mathfrak{X} \longrightarrow \mathcal{Y}_{T_n}$ una statistica (stimatore). Vi sia inoltre una funzione di parametri $a : \Theta \longrightarrow \mathcal{Y}_\Theta$. Allora la funzione non negativa $\mathrm{Loss} : (\mathcal{Y}_{T_n} \cup \mathcal{Y}_\Theta) \times \mathcal{Y}_\Theta \longrightarrow \mathbb{R}_{\ge 0}$ viene detta *Funzione di Perdita* se soddisfa alle seguenti condizioni:

1. $\mathrm{Loss}\big(a(\theta), a(\theta)\big) = 0$, $\forall\theta\in\Theta$;
2. Per ogni $T_n \in \mathcal{T}$, esiste una funzione $\mathrm{Risk} : Y_{T_n} \times Y_\Theta \longrightarrow \mathbb{R}$, detta *Funzione di Rischio*, tale che $\mathrm{Risk}\big(T_n, a(\theta)\big) = \mathbb{E}_\theta[\mathrm{Loss}\big(T_n, a(\theta)\big)]$, $\forall\theta\in\Theta$.

Osservazione. La funzione di perdita può essere pensata come una misura della discrepanza tra l'azione T_n e lo stato della natura $a(\theta)$.

Definizione. Possiamo già definire due tipologie di funzioni di perdita che spesso vengono utilizzate in statistica:

1. $\mathrm{Loss}_1\big(T_n, a(\theta)\big) \coloneqq |T_n - a(\theta)|$, chiamata *Errore assoluto*;
2. $\mathrm{Loss}_2\big(T_n, a(\theta)\big) \coloneqq \big(T_n - a(\theta)\big)^2$. Essa ammette anche come possibile funzione di rischio $\mathrm{Risk}_2\big(T_n, a(\theta)\big) \coloneqq \mathbb{E}_\theta\left[(T_n - a(\theta))^2\right]$; se tuttavia $a = \mathrm{id}_\Theta$, allora la funzione $\mathrm{MSE}_\theta(T_n) \coloneqq \mathrm{Risk}_2\big(T_n, \theta\big)$ prende il nome di *Mean Square Error* (oppure *Errore Quadratico Medio*).

Osservazione. Semplicemente aggiungendo e sottraendo il valore $\mathbb{E}[T_n]^2$ si ottiene subito la seguente uguaglianza: $\mathrm{MSE}_\theta(T_n) = \mathrm{Var}_\theta[T_n] + B_\theta[T_n]^2$.

Teorema 14. *Sia T_n uno stimatore di θ (non necessariamente non distorto). Allora si ha che $\lim_{n\to+\infty}\mathrm{MSE}_\theta(T_n) = 0$ è condizione sufficiente (ma non necessaria) per la consistenza di T_n.*

Dimostrazione. Si ha infatti la seguente semplice catena di diseguaglianze:

$$\begin{aligned}\mathbb{P}[|T_n-\theta|>\varepsilon] &= \int_{|T_n-\theta|>\varepsilon} f_{T_n}(\theta,t_n)dt_n \\ &< \int_{|T_n-\theta|>\varepsilon} \frac{(t_n-\theta)^2}{\varepsilon^2} f_{T_n}(\theta,t_n)dt_n < \frac{1}{\varepsilon^2}\,\mathrm{MSE}_\theta(T_n).\end{aligned}$$

□

1.4.2 Intervalli di confidenza

Lezione del 15/03, ultima modifica 26/03, Michele Nardin

Sia $(X_1, ..., X_n)$ un campione casuale definito da una variabile casuale avente funzione di ripartizione $F_X(x, \vartheta)$. Vogliamo stimare l'incognita ϑ, e per farlo ci serviamo di uno stimatore T_n. Una volta estratto il campione casuale, e quindi in possesso di una n-upla di valori reali $(x_1, ..., x_n)$ che ne rappresenta una determinazione, possiamo effettivamente calcolare valore della nostra stima: è impensabile però che la stima *coincida esattamente* con il valore incognito (se X ha distribuzione continua $\mathbb{P}(T_n = \vartheta) = 0$!). Dobbiamo quindi associare a T_n un *margine di errore*.

Introduciamo innanzitutto il concetto di Statistica Pivot:

Definizione 6. Sia $(X_1, ..., X_n)$ un campione casuale da una distribuzione con funzione di ripartizione $F_X(x, \vartheta)$, $\vartheta \in \Theta$. Definiamo Statistica Pivot una funzione $Q((X_1, ..., X_n), \vartheta)$ tale che

1. Q è funzione del campione casuale e del parametro ϑ (parametro su cui si vuol fare inferenza)
2. Q non contiene parametri incogniti oltre a ϑ
3. la distribuzione di Q, F_Q, è completamente nota (ossia non dipende da ϑ)
4. Q è invertibile rispetto a ϑ

Esempi: Campione casuale da $N(\mu, \sigma^2)$:

1. Supponiamo di conoscere la varianza: allora un esempio di statistica pivot è
$$Z_n = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}$$
la quale, grazie all'ipotesi di campionamento da vc normale, ha distribuzione N(0,1) (che non dipende da μ).
2. Supponiamo di non conoscere la varianza: in tal caso, al posto della varianza usiamo lo stimatore varianza campionaria S_n^2, il quale è non distorto (già dimostrato) e consistente (infatti $\mathrm{MSE}_{\sigma^2}(S_n^2) = Var(S_n^2) + B^2(S_n^2) = \frac{2\sigma^4}{n-1} \to 0$) e quindi la statistica pivot in questione sarà
$$Q = \frac{\bar{X}_n - \mu}{\S_n/\sqrt{n}}$$
la quale (dimostreremo che) ha distribuzione t-student con n-1 gradi di libertà.

Esempio introduttivo (exit poll)

Vogliamo stimare la proporzione p_i dei voti ricevuti dall'iesimo partito sul totale. Il nostro problema sarà quello di trovare un intervallo centrato nella stima $\hat{p_i}$, ed un margine d'errore, ME, tale per cui, ad una fissata soglia di probabilità α si abbia

$$\mathbb{P}[p_i \in (\hat{p_i} - ME, \hat{p_i} + ME)] = 1 - \alpha$$

Costruzione generale

In generale, sia ϑ_0 il valore vero del parametro ϑ che vogliamo stimare, e per semplicità assumiamo che T_n sia un suo stimatore tale che

$$\sqrt{n}(T_n - \vartheta_0) \xrightarrow{d} N(0, \sigma^2_{T_n})$$

Per il momento assumiamo di conoscere $\sigma^2_{T_n}$, sicché

$$Z_n = \frac{\sqrt{n}(T_n - \vartheta_0)}{\sigma_{T_n}} \overset{a}{\sim} N(0,1)$$

Bisogna notare che Z_n è una statistica pivot. Fissato $\alpha \in (0,1)$, consideriamo i quantili della distribuzione N(0,1), $\pm z_{\alpha/2}$ (ossia quei valori tali per cui, se $X \sim N(0,1)$, $P(-z_{\alpha/2} \leq X \leq z_{\alpha/2}) = 1-\alpha$). Possiamo affermare che, per n sufficientemente grande, (il simbolo $\doteq$ indica un'uguaglianza approssimata)

$$P(-z_{\alpha/2} \leq Z_n \leq z_{\alpha/2}) \doteq 1 - \alpha$$

da cui

$$P(-z_{\alpha/2} \leq \frac{\sqrt{n}(T_n - \vartheta_0)}{\sigma_{T_n}} \leq z_{\alpha/2}) \doteq 1 - \alpha$$

e ancora

$$P(T_n - z_{\alpha/2}\frac{\sigma_{T_n}}{\sqrt{n}} \leq \vartheta_0 \leq T_n + z_{\alpha/2}\frac{\sigma_{T_n}}{\sqrt{n}}) \doteq 1 - \alpha$$

Possiamo quindi definire un intervallo casuale,

$$IC = \left[T_n - z_{\alpha/2}\frac{\sigma_{T_n}}{\sqrt{n}}, T_n + z_{\alpha/2}\frac{\sigma_{T_n}}{\sqrt{n}}\right]$$

(è casuale perchè per T_n è una vc). IC è uno Stimatore Intervallare. Si può affermare che $P(\vartheta \in IC) \doteq 1 - \alpha$.

Nomenclatura : $z_{\alpha/2}$ si dice Fattore di Affidabilità, $\frac{\sigma_{T_n}}{\sqrt{n}}$ si dice Standard Error dello stimatore T_n.

Sia ora $(x_1, ..., x_n)$ una determinazione campionaria (ossia i dati effettivamente osservati da un campione casuale) (cioè una n-upla) e sia $T_n(x_1, ..., x_n) = t_n$ l'effettivo valore assunto dallo stimatore.

Definiamo di seguito *l'intervallo di confidenza con probabilità di copertura* $1 - \alpha$

$$IC_\vartheta(1-\alpha) := \left[t_n - z_{\alpha/2}\frac{\sigma_{T_n}}{\sqrt{n}}, t_n + z_{\alpha/2}\frac{\sigma_{T_n}}{\sqrt{n}}\right]$$

La probabilità di copertura viene anche detta livello di confidenza.

Nella pratica, $\sigma^2_{T_n}$ non è noto a priori. Possiamo però usare lo stimatore varianza campionaria di T_n, $S^2_{T_n}$, il quale sappiamo che converge in probabilità a $\sigma^2_{T_n}$. Allora, per il teorema 10 (di Slutsky), troviamo che

$$Z_n = \frac{\sqrt{n}(T_n - \vartheta_0)}{S_{T_n}} = \frac{\sqrt{n}}{S_{T_n}} T_n - \frac{\sqrt{n}}{S_{T_n}} \vartheta_0 \xrightarrow{d} N(0,1)$$

Possiamo quindi ripetere il ragionamento fatto poco sopra usando la varianza campionaria al posto di $S^2_{T_n}$, e quindi costruire l'intervallo di confidenza con probabilità di copertura pari a $1-\alpha$ come

$$IC_\vartheta(1-\alpha) := \left[t_n - z_{\alpha/2} \frac{S_{T_n}}{\sqrt{n}}, t_n + z_{\alpha/2} \frac{S_{T_n}}{\sqrt{n}} \right]$$

Intervallo di confidenza per la media μ

Sia $(X_1, ..., X_n)$ un campione casuale, media e varianza incognite. Siano $\bar{X}_n$ e S^2_n gli stimatori di media e varianza della popolazione. Allora per il TLC e per il teorema di Slutsky si ha che

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{S_n} \xrightarrow{d} N(0,1)$$

che è una statistica pivot. Quindi l'intervallo di confidenza con probabilità di copertura $1-\alpha$ (sempre approssimato) sarà

$$IC_\mu(1-\alpha) = \left[\bar{X}_n - z_{\alpha/2} \frac{S_n}{\sqrt{n}}, \bar{X}_n + z_{\alpha/2} \frac{S_n}{\sqrt{n}} \right]$$

Intervallo di confidenza per una proporzione p

Sia $(X_1, ..., X_n)$ un campione casuale da $b(1,p)$ e sia $\hat{p}_n = \frac{1}{n}\sum_{i=1}^n X_i$ lo stimatore (corretto e consistente) di p. Troviamo che per il TLC e per la WLLN (legge dei grandi numeri)

$$\frac{\sqrt{n}(\hat{p}_n - p)}{\sqrt{\hat{p}_n(1-\hat{p}_n)}} \xrightarrow{d} N(0,1)$$

e quindi l'intervallo di confidenza con probabilità di copertura $1-\alpha$ approssimato sarà

$$IC_p(1-\alpha) = \left[\hat{p}_n - z_{\alpha/2} \sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n}}, \hat{p}_n + z_{\alpha/2} \sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n}} \right]$$

Distribuzione esatta della statistica pivot: distribuzione t di Student

Lezione del 18/03, ultima modifica 26/03, Michele Nardin

La distribuzione t di Student con ν gradi di libertà è definita come $T = \frac{Z}{\sqrt{S^2/\nu}}$ ove $Z \sim N(0,1)$ mentre $S^2 \sim \chi^2_\nu$ (chiquadro con ν gradi di libertà). La funzione di densità è

$$f_{t_\nu}(t,\nu) = \frac{\Gamma((\nu+1)/2)}{\Gamma(\nu/2)} \frac{1}{\sqrt{\pi\nu}} \frac{1}{[1+t^2/\nu]^{\frac{v+1}{2}}} \mathbb{1}_{\mathbb{R}}(t)$$

tale funzione è simmetrica, ha la classica forma a campana come la normale, ma a differenza di quest'ultima ha le code più pesanti. Risulta che la statistica pivot per la media in campioni poco numerosi [1] (in caso di campionamento da normale) ha distribuzione esatta t di Student. Infatti

$$Q = \frac{\overline{X}_n - \mu}{S_n/\sqrt{n}} = \frac{\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}}}{\sqrt{\frac{S_n^2}{\sigma^2}}}$$

troviamo al numeratore $\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$, (grazie al fatto che le X_i sono equi distribuite normalmente) mentre al denominatore abbiamo che

$$\sqrt{\frac{S_n^2}{\sigma^2}} = \sqrt{\frac{(n-1)S_n^2}{(n-1)\sigma^2}} = \sqrt{\frac{H}{(n-1)}}$$

Abbiamo già dimostrato che $H = \frac{(n-1)S_n^2}{\sigma^2} \sim \chi^2_{n-1}$, quindi in definitiva al denominatore abbiamo la radice di una chiquadro diviso i suoi gradi di libertà, ovvero siamo proprio in presenza di una distribuzione t di Student.

Osservazione importante: Quindi, quando il campione casuale è poco numeroso, è conveniente usare i quantili della distribuzione t di student per costruire gli intervalli di confidenza. Per numerosità campionarie $n > 30$, approssimare la distribuzione t di student con la distribuzione normale offre risultati soddisfacenti. Ricordiamo che per il tlc $Q \to N(0,1)$)

Intervallo di confidenza esatto

Fissato un livello di confidenza $1-\alpha$, consideriamo i quantili della distribuzione t di student (con n-1 gradi di libertà, ove n è la dimensione campionaria) $\pm t_{(\alpha/2;n-1)}$, troviamo

$$P\left(-t_{(\alpha/2;n-1)} \leq \frac{\overline{X}_n - \mu}{S_n/\sqrt{n}} \leq t_{(\alpha/2;n-1)}\right) = 1-\alpha$$

Notiamo che questa volta vale l'uguaglianza 'vera', poiché non stiamo considerando approssimazioni asintotiche. In presenza del campione effettivamente estratto, $(x_1, ..., x_n)$, scriviamo $\overline{x}_n$ e s_n^2 i valori assunti da media e varianza campionaria, l'intervallo di confidenza è

$$IC_\mu(1-\alpha) = \left[\overline{x}_n - t_{(\alpha/2;n-1)}\sqrt{\frac{s_n^2}{n}}, \overline{x}_n + t_{(\alpha/2;n-1)}\sqrt{\frac{s_n^2}{n}}\right]$$

Osservazione 7. Alcune osservazioni che, pur sembrando banali, è bene tenere a mente:

1. Al crescere del livello di confidenza $(1-\alpha)$ e/o della varianza campionaria S_n^2 cresce anche l'ampiezza di IC
2. Al crescere dell'ampiezza campionaria n, (fermo restando il livello di confidenza) l'ampiezza di IC diminuisce

[1] In realtà vale per tutti i campioni, è solo che da un certo punto in poi la differenza con la normale è davvero trascurabile! Sulle tavole si riporta solo per $\nu < 120$

Intervalli di confidenza per la varianza

Sia $(X_1, ..., X_n)$ un campione casuale da $N(\mu, \sigma^2)$. Consideriamo la statistica pivot

$$W = \frac{n-1}{\sigma^2} S_n^2$$

Abbiamo già mostrato che $W \sim \chi^2_{n-1}$. Ma allora, dato che noi cerchiamo q_1, q_2 t.c.

$$P\left(q_1 \leq \frac{n-1}{\sigma^2} S_n^2 \leq q_2\right) = 1 - \alpha$$

troviamo che essi sono i quantili di ordine $\alpha/2$ e $1 - \alpha/2$ della chiquadro con n-1 gradi di libertà, che indicheremo $q_1 = \chi^2_{(n-1,\alpha/2)}$ e $q_2 = \chi^2_{(n-1,1-\alpha/2)}$. Con qualche passaggio otteniamo:

$$P\left(\frac{1}{q_2} \leq \frac{\sigma^2}{(n-1)S_n^2} \leq \frac{1}{q_1}\right) = 1 - \alpha$$

$$P\left(\frac{(n-1)S_n^2}{q_2} \leq \sigma^2 \leq \frac{(n-1)S_n^2}{q_1}\right) = 1 - \alpha$$

Troviamo così l'intervallo casuale (e di conseguenza il relativo intervallo di confidenza, una volta estratto il campione e trovato un valore a S_n^2)

$$IC = \left[\frac{(n-1)S_n^2}{q_2}; \frac{(n-1)S_n^2}{q_1}\right]$$

Intervalli di confidenza per la differenza di medie

Vogliamo confrontare due distribuzioni: *sintetizziamo* la differenza tra due popolazioni tramite la differenza delle loro media.

Supponiamo inizialmente di avere due campioni casuali tra loro indipendenti:
$(X_1, ..., X_{n_1})$ da una distribuzione D1, con media μ_1 (ignota) e varianza σ_1^2 (nota)
$(Y_1, ..., Y_{n_2})$ da una distribuzione D2, con media μ_2 (ignota) e varianza σ_2^2 (nota)
NB: non necessariamente n_1 dev'essere uguale a n_2

Consideriamo gli stimatori media campionaria per le due medie, che indicheremo con $\overline{X}$ e $\overline{Y}$. La statistica pivot che ci interessa per $\Delta = \mu_1 - \mu_2$ sarà

$$Z = \frac{(\overline{X} - \overline{Y}) - (\mu_1 - \mu_2)}{\left[\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right]^{\frac{1}{2}}}$$

Notiamo che $var(\overline{X} - \overline{Y}) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$ dato che $cov(\overline{X}, \overline{Y}) = 0$ per l'indipendenza. Ma allora $Z \overset{a}{\sim} N(0,1)$, quindi possiamo trovare un intervallo di confidenza [2]

$$IC_\Delta(1-\alpha) = [(\overline{X} - \overline{Y}) - ME; (\overline{X} - \overline{Y}) + ME]$$

[2] Approssimato, visto che conosciamo solo l'andamento asintotico di Z! D1 e D2 non è detto che siano mormali!

ove $ME = z_{\alpha/2}\left[\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right]^{\frac{1}{2}}$. Al posto delle varianze possiamo usare anche gli stimatori corretti e consistenti varianza campionaria, e giungere allo stesso risultato per il teorema di Slutsky.

In generale non conosciamo la varianza delle distribuzioni: in base al problema che dobbiamo affrontare, può essere plausibile supporre di conoscere la distribuzione delle due popolazioni a meno di uno o più parametri.

Location Model: Supponiamo di avere $(X_1, ..., X_{n_1})$ da distribuzione normale con media μ_1 e varianza σ_1^2 (ignote), i loro stimatori $\overline{X}$ e S_1^2 e $(Y_1, ..., Y_{n_2})$ da distribuzione normale con media μ_2 e varianza σ_2^2 (ignote) e i loro stimatori $\overline{Y}$ e S_2^2. Supponiamo che i due campioni siano tra loro indipendenti ed inoltre che $\sigma_1 = \sigma_2 = \sigma$. Possiamo 'fondere' le informazioni contenute in S_1^2 e S_2^2:

$$(PooledVariance)\ S_p^2 := \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

che risulta essere uno stimatore corretto e consistente di σ^2 (esercizio). La statistica Pivot che prendiamo in considerazione sarà

$$T = \frac{(\overline{X} - \overline{Y}) - (\mu_1 - \mu_2)}{S_p\left(\frac{1}{n_1} + \frac{1}{n_2}\right)^{\frac{1}{2}}}$$

la quale risulta essere distribuita come $t_{n_1+n_2-2}$. Ricalcando i passaggi delle applicazioni precedenti, fissato α troviamo l'intervallo casuale per Δ

$$IC = \left[(\overline{X} - \overline{Y}) - t_{(n_1+n_2-2;\alpha/2)}S_p\left(\frac{1}{n_1} + \frac{1}{n_2}\right)^{\frac{1}{2}}; (\overline{X} - \overline{Y}) + t_{(n_1+n_2-2;\alpha/2)}S_p\left(\frac{1}{n_1} + \frac{1}{n_2}\right)^{\frac{1}{2}}\right]$$

Intervalli di confidenza per la differenza di proporzioni

Supponiamo di avere $(X_1, ..., X_{n_1})$ da distribuzione $b(1, p_1)$, con stimatore $\hat{p_1}$ e $(Y_1, ..., Y_{n_2})$ da distribuzione $b(1, p_2)$, con stimatore $\hat{p_2}$. Supponiamo che i due campioni siano tra loro indipendenti. Allora

$$\Delta = \hat{p_1} - \hat{p_2} \overset{a}{\sim} N\left(p_1 - p_2, \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}\right)$$

quindi usando la statistica Pivot

$$Z = \frac{(\hat{p_1} - \hat{p_2}) - (p_1 - p_2)}{\sqrt{\left(p_1 - p_2, \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}\right)}} \overset{a}{\sim} N(0, 1)$$

trovo l'intervallo di confidenza

$$IC_\Delta(1 - \alpha) = \left[(\hat{p_1} - \hat{p_2}) - z_{\alpha/2}\sqrt{A(p_1, p_2)}; (\hat{p_1} - \hat{p_2}) + z_{\alpha/2}\sqrt{A(p_1, p_2)}\right]$$

ove $A(p_1, p_2) = \left(\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}\right)$. Ovviamente al posto di p_1 e p_2 uso gli stimatori corretti e consistenti $\hat{p_1}$ e $\hat{p_2}$.

Lezione del 20 marzo, ultima modifica 26 marzo, Michele Nardin

Intervalli di confidenza per rapporti di varianze

Introduciamo per prima cosa la *Distribuzione F di Snedecor-Fisher*.

Siano $W_1 \sim \chi^2_{\nu_1}$ e $W_2 \sim \chi^2_{\nu_2}$ indipendenti. La distribuzione F_{ν_1,ν_2} è definita come il rapporto tra due chiquadrato divise per i rispettivi gradi di libertà, in formule

$$W = \frac{W_1/\nu_1}{W_2/\nu_2}$$

ed ha funzione di densità

$$f_W(w;\nu_1,\nu_2) = \frac{\Gamma(\frac{\nu_1+\nu_2}{2})}{\Gamma(\frac{\nu_1}{2}) + \Gamma(\frac{\nu_2}{2})}(\nu_1/\nu_2)^{\nu_1/2} w^{\nu_1/2-1}\left[1 - \frac{\nu_1}{\nu_2}w\right]^{\frac{\nu_1+\nu_2}{2}} \mathbb{1}_{\mathbb{R}^+}(w)$$

Vale la seguente proprietà, utile per calcolare i quantili non tabulati:

$$Se\ W \sim F_{\nu_1,\nu_2} \Rightarrow \frac{1}{W} \sim F_{\nu_2,\nu_1}$$

Le tavole (comunemente) forniscono i valori dei quantili per $(1-\alpha) \in \{0.80, 0.90, 0.95, 0.975, 0.99, 0.999\}$. Quindi possiamo sfruttare la proprietà sopra scritta per trovare che

$$w_{\alpha;\nu_1,\nu_2} = \frac{1}{w_{1-\alpha;\nu_2,\nu_1}}$$

Intervallo di confidenza per rapporti di varianze

Supponiamo di avere $(X_1, ..., X_{n_1})$ da distribuzione normale con media μ_1 e varianza σ_1^2 (ignote), i loro stimatori $\overline{X}$ e S_1^2 e $(Y_1, ..., Y_{n_2})$ da distribuzione normale con media μ_2 e varianza σ_2^2 (ignote) e i loro stimatori $\overline{Y}$ e S_2^2. Ricordiamo che $\frac{(n_1-1)S_1^2}{\sigma_1^2} \sim \chi^2_{n_1-1}$ (idem per S_2^2).

Fissato $1-\alpha$ consideriamo una statistica pivot e w_1, w_2 tc $P(w_1 \leq W \leq w_2) = 1-\alpha$. La statistica pivot in questione sarà

$$W = \frac{\frac{(n_1-1)S_1^2}{\sigma_1^2}/n_1 - 1}{\frac{(n_2-1)S_2^2}{\sigma_2^2}/n_2 - 1} \sim F_{(n_1-1),(n_2-1)}$$

poichè rapporto di due chiquadro. Risulta inoltre, semplificando:

$$W = \frac{S_1^2}{\sigma_1^2}\frac{S_2^2}{\sigma_2^2} = \frac{\sigma_2^2}{\sigma_1^2}\frac{S_1^2}{S_2^2}$$

w_1 e w_2 saranno i quantili di ordine $\alpha/2$ e $1-\alpha/2$ della distribuzione $F_{(n_1-1),(n_2-1)}$, esplicitamente $w_1 = w_{(n_1-1,n_2-1;\alpha/2)} = \frac{1}{w_{(n_2-1,n_1-1;1-\alpha/2)}}$ e $w_2 = w_{(n_1-1,n_2-1;1-\alpha/2)}$ e quindi troviamo che

$$1-\alpha = P(w_1 \leq \frac{\sigma_2^2}{\sigma_1^2}\frac{S_1^2}{S_2^2} \leq w_2) = P\left(\frac{S_1^2/S_2^2}{w_2} \leq \frac{\sigma_1^2}{\sigma_2^2} \leq \frac{S_1^2/S_2^2}{w_1}\right)$$

da cui troviamo anche il relativo intervallo casuale. Quando invece abbiamo un campione effettivamente estratto, posti s_1^2 e s_2^2 i valori assunti dalla varianza campionaria, l'intervallo di confidenza sarà

$$IC_{\frac{\sigma_1^2}{\sigma_2^2}}(1-\alpha) = \left[\frac{s_1^2/s_2^2}{w_2}; \frac{s_1^2/s_2^2}{w_1}\right] = \left[\frac{s_1^2/s_2^2}{w_{(n_1-1,n_2-1;1-\alpha/2)}}; \frac{s_1^2/s_2^2}{\frac{1}{w_{(n_2-1,n_1-1;1-\alpha/2)}}}\right]$$

1.5 Test di ipotesi

Lezione del 25/03, ultima modifica 20/05, Andrea Gadotti

La procedura di test per la verifica di ipotesi che descriveremo a breve cerca di fornire una soluzione ai seguenti problemi:

1. Determinare quanto un'ipotesi è realistica, verosimile, compatibile con l'informazione empirica a disposizione.
2. Trovare un ragionamento oggettivo (matematico) per inferire dall'informazione disponibile (ovvero il contenuto di un campione) circa la veridicità dell'ipotesi formulata.
3. Misurare in qualche modo questa "vicinanza" tra ipotesi e realtà.

Useremo statistiche pivot in ambito parametrico: la distribuzione da cui proviene il campione casuale $(X_1, ..., X_n)$ è nota a meno di uno o più parametri.

1.5.1 Tipi di ipotesi

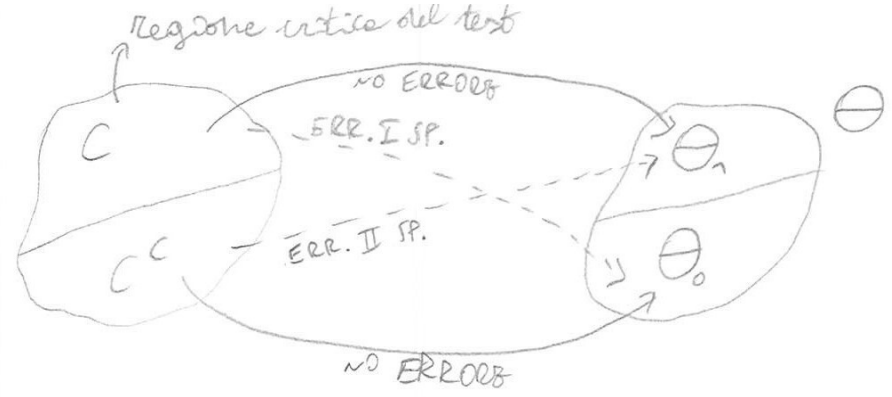
Supponiamo di avere un campione casuale i.i.d. $(X_1, \ldots, X_n) \sim F_X(\mathbf{x}; \theta)$. Supponiamo che lo spazio Θ in cui vive il parametro θ sia partizionato in due sottoinsiemi Θ_0, Θ_1. Le *ipotesi* vengono poste nella forma:

$$\begin{cases} H_0 \colon \theta \in \Theta_0 \\ H_1 \colon \theta \in \Theta_1, \end{cases} \quad \Theta = \Theta_0 \cup \Theta_1. \tag{1.8}$$

Chiameremo H_0 *ipotesi nulla* e H_1 *ipotesi alternativa*. Idealmente, H_0 rappresenta la conoscenza pregressa, la supposizione vera fino a prova contraria; invece, H_1 costituisce l'ipotesi di lavoro, quella su cui ripieghiamo nel momento in cui il nostro test risulta in contraddizione con H_0.

Il test si riduce a una *regola di decisione* in merito a H_0 e H_1 sulla base del campione casuale $(X_1, ..., X_n)$ da $X \sim F_X(x; \theta)$. Dividiamo lo spazio dei campioni in due regioni disgiunte: C (regione critica del test) e C^c. La decisione può chiaramente essere corretta, ma anche errata, poiché il campione costituisce un'informazione non completa. Risulta quindi necessario formulare delle *conclusioni in probabilità*, ovvero associare alla nostra conclusione la probabilità che questa sia corretta, cercando ovviamente di massimizzarla. Possiamo riassumere le varie possibilità nella tabella e nel disegno sottostanti:

	H_0 è **vera**	H_0 è **falsa**
Rifiuto H_0	errore di specie I	nessun errore
Non rifiuto H_0	nessun errore	errore di specie II



Lancio di una moneta. Consideriamo il campione casuale $(X_1, ..., X_n) \sim b(1, p)$ rappresentante n lanci di una moneta. Vogliamo testare l'onestà della moneta, in particolare capire se essa è truccata per far uscire più frequentemente croce. In questo caso, ipotizzeremo:

$$\begin{cases} H_0 : p \geq \frac{1}{2} \\ H_1 : p < \frac{1}{2}. \end{cases}$$

e il numero di teste $S_n = \sum_{i=1}^{n} X_i$. Vorremmo stimare la probabilità che esca testa con la media campionaria $\overline{X}_n$. In questo caso potremmo avere:

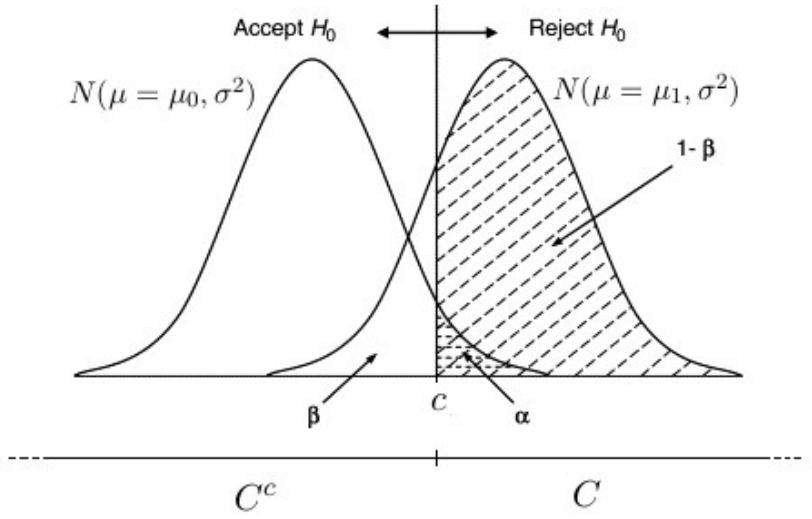
$$\begin{cases} H_0 : & p = 1/2 \\ H_1 : & p \neq 1/2 \end{cases}$$

La regola di decisione consiste quindi nel rifiutare H_0 se $(X_1, ..., X_n) \in C$ e invece rifiutare H_0 se $(X_1, ..., X_n) \in C^c$. Ci piacerebbe trovare una regola di decisione che permetta di minimizzare la probabilità di commettere errori di I o II tipo. Purtroppo questo non è possibile, per la natura stessa della relazione che corre tra gli errori di I e II tipo. Di seguito un esempio che ci dà un'idea del perché:

Esempio Consideriamo un campione casuale $(X_1, ..., X_n)$ da $N(\mu, \sigma^2)$ con σ^2 noto. Supponiamo che le nostre due ipotesi siano:

$$\begin{cases} H_0: & \mu = \mu_0 \quad \text{ovvero } N(\mu = \mu_0, \sigma^2) \\ H_1: & \mu = \mu_1 \quad \text{ovvero } N(\mu = \mu_1, \sigma^2) \end{cases}$$

con $\mu_1 > \mu_0$.



Consideriamo

$$\begin{aligned} \alpha :&= \mathbb{P}(\text{rifiutare } H_0 \mid H_0 \text{ vera}) \\ &= \mathbb{P}(\text{campione } \in C \mid H_0 \text{ vera}) \\ &= \mathbb{P}(\text{il nostro campione è } \geq c \mid \text{la distribuzione corretta è quella di sinistra}) \\ &= P(\text{commettere un errore di I tipo}). \end{aligned}$$

e

$$\begin{aligned} \beta :&= P(\text{non rifiutare } H_0 \mid H_0 \text{ falsa}) \\ &= P(\text{il nostro campione appartiene a } C^c \mid H_0 \text{ falsa}) \\ &= P(\text{il nostro campione è } \leq c \mid \text{la distribuzione corretta è quella di destra}) \\ &= P(\text{commettere un errore di II tipo}) \end{aligned}$$

. (Nota: α è detto *livello di significatività del test*)
È evidente che non è possibile annullare contemporaneamente sia α che β.
La procedura si divide quindi in due passi: il primo consiste nel **fissare** α, il secondo nell'individuare la regola di decisione che minimizza β, in modo da trovare un test *ottimo*.

Lezione del 05/04, ultima modifica 20/05, Andrea Gadotti

In generale una statistica test si può descrivere come di seguito:

$$\begin{cases} H_0: & \theta \in \Theta_0 \\ H_1: & \theta \in \Theta_1 \end{cases}$$

dove Θ è lo spazio dei possibili parametri della distribuzione e $\Theta = \Theta_0 \cup \Theta_1$.
Quello che vogliamo trovare è la regola di partizionamento che divida lo spazio dei campioni C in C_0 e C_1 in funzione di α (deciso da noi). Per farlo imponiamo la condizione $\alpha = P(\underline{x} \in C \mid \theta \in \Theta_0)$. Vediamo ora un esempio con il lancio di una moneta:

Esempio Consideriamo il campione casuale $(X_1, ..., X_n)$ dove $X_i = 0$ con probabilità p e $X_i = 1$ con probabilità $1 - p$. Facciamo le nostre ipotesi:

$$\begin{cases} H_0: & p \leq 1/2 \\ H_1: & p > 1/2 \end{cases}$$

Prediamo ora come regola di decisione $\frac{S}{n} = \frac{\sum X_i}{n}$. Deciso un α a nostra discrezione, imponiamo l'equazione in k:

$$\alpha = P(S > k \mid p \leq 1/2) \ (= P(S > k \mid H_0 \text{ vera}))$$

A questo punto risolvendo l'equazione troveremo il k per il quale rifiuteremo H_0 se $S > k$.

Esempio Sia $(X_1, ..., X_n)$ un campione casuale con $X_i \sim b(1, p)$ e sia

$$\begin{cases} H_0: & p = p_0 \\ H_1: & p < p_0 \end{cases}$$

Sulla base dell'informazione circa p contenuta in $(X_1, ..., X_n)$ vogliamo sottoporre a verifica il sistema di ipotesi in questione. Procediamo in questo ordine:

(a) Prendiamo $S := \sum X_i \sim b(n, p)$, che è di fatto il numero di successi. Sotto H_0 abbiamo che $S \sim b(n, p_0)$.

(b) Scegliamo una regola di decisione (usando anche la distribuzione -nota- di S sotto H_0). Ovvero, individuiamo la *regione di rifiuto del test*. A questo punto vorremo rifiutare H_0 a favore di H_1 quando $S \leq k$, dove k è l'incognita che troveremo nel punto (c).

(c) Scelto il nostro α, si ha che il valore di k deve essere tale per cui

$$\alpha = P(S \leq k \mid p = p_0) = \sum_{s=0}^{k} \binom{n}{s} p_0^s (1 - p_0)^{n-s}$$

A questo punto, essendo α fissato, abbiamo un'equazione in k che risolta ci restituisce il suo valore.

Esempio particolare Nella situazione generale sopra descritta prendiamo un caso particolare con $n = 20$ e $p_0 = 0,7$. Decidiamo $\alpha = 0.15$. L'equazione diventa: $0,15 = \sum_{s=0}^{k} \binom{20}{s} 0,7^s 0,3^{20-s}$. Osserviamo che il valore di $P(S \leq k \mid p = 0,7)$ per $k = 11$ risulta $0,1133$, mentre per $k = 12$ è $0,2277$. Quindi il nostro k è compreso tra 11 e 12. In conclusione, se il nostro test dovesse presentare 12 (o più) successi, allora non rifiuteremmo H_0. In caso contrario scarteremmo H_0 a favore di H_1.

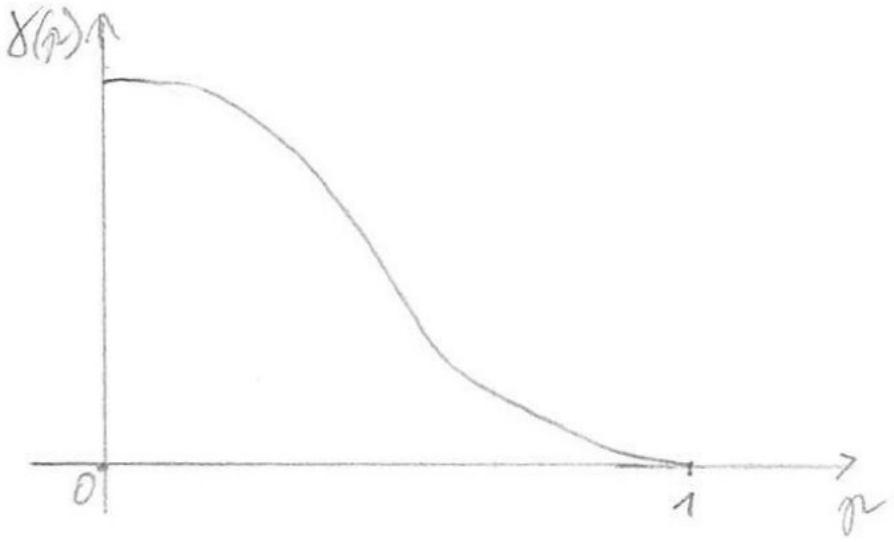
Definizione 7. Sia $\beta := P(\underline{x} \in C_0 \mid \theta \in \Theta_1)$, ovvero β è la probabilità (fissato α) di commettere un errore di II specie. Chiamiamo *potenza del test* il valore $\gamma := 1 - \beta$. Un test risulta ottimale quando la sua potenza è massima. Notiamo che possiamo definire una *funzione di potenza* $\gamma(t) := 1 - \beta(t)$

Osservazione 8. Prendendo di nuovo in considerazione l'esempio precedente sul campione casuale normale, è chiaro che una volta fissato α, ovvero c, minore è μ_1, più piccola è l'area sottesa dalla coda della relativa normale, ovvero β.

Esempio Prendendo di nuovo in considerazione l'esempio precedente sulla bernoulliana, abbiamo che

$$\gamma(p) = 1 - \beta(p) = 1 - P(S > k \mid p < p_0) = P(S \leq k \mid p < p_0) = \sum_{s=0}^{k} \binom{n}{s} p^s (1-p)^{n-s}$$

Di seguito possiamo osservare il grafico della funzione:



Osservazione Il test in merito al precedente sistema di ipotesi relative a p è un *test esatto*, perché poggia sulla distribuzione *esatta* di S ($S \sim b(n,p)$). Questo però non accade sempre, e quindi talvolta è necessario ricorrere alla teoria asintotica e dei test approssimati.

1.5.2 Esempi di statistiche test (generali e particolari)

Test per la media di una popolazione qualsiasi Supponiamo di avere un campione casuale $(X_1, ..., X_n)$ proveniente da una distribuzione non nota di media μ e varianza σ^2 (finita) non note.
Le nostre ipotesi sono

$$\begin{cases} H_0: & \mu = \mu_0 \\ H_1: & \mu = \mu_1 \quad \text{con } \mu_1 > \mu_0 \end{cases}$$

Decidiamo di "condensare" l'informazione presente nel campione circa μ e σ^2 tramite $\overline{X}_n$ e S_n^2 (che ricordiamo essere stimatori non distorti e consistenti), sapendo che $\overline{X}_n \xrightarrow{P} \mu$ e $S_n^2 \xrightarrow{P} \sigma^2$.
A questo punto, la nostra regola di decisione consisterà nel rifiutare H_0 a favore di H_1 se $\overline{X}_n$ è molto più grande di μ_0.
Noi sappiamo che $\overline{X}_n \overset{a}{\sim} N\left(\mu, \frac{S_n^2}{n}\right)$, ovvero $\frac{\overline{X}_n - \mu}{S_n/\sqrt{n}} \xrightarrow{D} N(0,1) =: Z$
Usando questo risultato possiamo individuare la regione critica del test di livello α fissato. Imponiamo la seguente uguaglianza:

$$\begin{aligned} \alpha &= P(\overline{x} \in C \mid \mu = \mu_0) = P(\overline{X}_n \geq k \mid \mu = \mu_0) \\ &= P\left(\frac{\overline{X}_n - \mu}{S_n/\sqrt{n}} \geq \frac{k - \mu}{S_n/\sqrt{n}} \mid \mu = \mu_0\right) \\ &= P\left(\frac{\overline{X}_n - \mu_0}{S_n/\sqrt{n}} \geq \frac{k - \mu_0}{S_n/\sqrt{n}}\right) \end{aligned}$$

Quindi

$$1-\alpha = 1-P\left(\frac{\overline{X}_n - \mu_0}{S_n/\sqrt{n}} \geq \frac{k - \mu_0}{S_n/\sqrt{n}}\right) = P\left(\frac{\overline{X}_n - \mu_0}{S_n/\sqrt{n}} < \frac{k - \mu_0}{S_n/\sqrt{n}}\right) = P\left(N(0,1) < \frac{k - \mu_0}{S_n/\sqrt{n}}\right)$$

ovvero

$$\frac{k - \mu_0}{S_n/\sqrt{n}} = z_{1-\alpha} = -z_\alpha$$

Dove $z_{1-\alpha}$ è il valore da cercare sulle tavole relative alla distribuzione normale in funzione dell'α scelto. Nota: l'ultima uguaglianza di probabilità è in realtà un'approssimazione che è tanto più corretta quanto più grande è n.
In definitiva, abbiamo che $C = \{\underline{x} \in \varkappa : \frac{\overline{X}_n - \mu_0}{S_n/\sqrt{n}} \geq -z_\alpha\} = \{\underline{x} \in \varkappa : \overline{X}_n \geq \mu_0 - z_\alpha \frac{S}{\sqrt{n}}\}$

Possiamo anche considerare la *funzione di potenza approssimata*:

$$\begin{aligned} \gamma(\mu_1) = 1 - \beta(\mu_1) &= P(\overline{X}_n \geq \mu_0 - z_\alpha \sigma/\sqrt{n} \mid \mu = \mu_1) \\ &= P\left(\frac{\overline{X}_n - \mu_1}{\sigma/\sqrt{n}} \geq \frac{\mu_0 - z_\alpha \sigma/\sqrt{n} - \mu_1}{\sigma/\sqrt{n}}\right) \\ &= 1 - P\left(Z \geq -z_\alpha + \frac{\sqrt{n}(\mu_0 - \mu_1)}{\sigma}\right) \\ &= \Phi\left(-z_\alpha + \frac{\sqrt{n}(\mu_0 - \mu_1)}{\sigma}\right) \end{aligned}$$

dove Φ è la funzione di ripartizione di $N(0,1)$.
Notiamo che il valore di $\gamma(\mu_1)$ tende a 1 per $n \to \infty$. Intuitivamente, questo è esattamente ciò che ci aspettiamo, in quanto più è grande μ_1, più esso è distante dal nostro μ_0, e di conseguenza è lecito aspettarsi che la probabilità che un campione abbia media vicina a μ_0 quando invece $\mu = \mu_1$ sarà bassa, ovvero la potenza del test è elevata.
È chiaro quindi che una funzione di potenza è tanto migliore quanto più il suo grafico sta vicino alla retta $y = 1$.

Esempio In riferimento al caso generale appena trattato, supponiamo di avere $\mu_0 = 12$, $\overline{X}_n = 14,3$, $S_n^2 = 22,5$, $n = 50$. Se fissiamo $\alpha = 0,05$, usando le tavole per la distribuzione normale $N(0,1)$ troviamo $z_\alpha = 1,645$. Ne segue che $k = 12+1,645\sqrt{22,5/50}$, che è minore di $14,3$. Concludiamo quindi rifiutando H_0.

Esempio di test esatto con t di Student Abbiamo $(X_1, ..., X_n)$ campione casuale da $N(\mu, \sigma^2)$ con μ e σ^2 non noti. Le nostre ipotesi sono:

$$\begin{cases} H_0 : & \mu = \mu_0 \\ H_1 : & \mu > \mu_0 \end{cases}$$

Sappiamo che $\overline{X}_n \overset{H_0}{\sim} N(\mu_0, \sigma^2/n)$ e quindi:

$$T := \frac{\overline{X}_n - \mu_0}{S_n/\sqrt{n}} = \frac{\overline{X}_n - \mu_0}{\sigma/\sqrt{n}} \frac{1}{\sqrt{S_n^2/\sigma^2}} \sim \frac{Z}{S_n/\sqrt{n}} \sim t_{n-1}$$

(vedi pag. 17 e pag. 12)
A questo punto, possiamo trovare il nostro valore critico k usando le tavole della distribuzione t_{n-1}.
Notiamo che quello appena mostrato è un *test esatto*, in quanto non si basa su un'approssimazione dello stimatore per valori elevati di n (usando ad esempio il TLC), bensì usa la sua distribuzione reale (in questo caso la distribuzione t_{n-1}).

Lezione 08/04, ultima modifica 20/05, Michele Nardin

Esempio di test bilaterale Sia $(X_1, ..., X_n)$ un campione casuale da una distribuzione avente media μ e varianza σ^2 finite. Considerato il fatto che non abbiamo informazioni sulla distribuzione delle variabili casuali, ci appoggeremo su di un test *approssimato*. Supponiamo di voler verificare la seguente ipotesi relativa alla media della distribuzione:

$$\begin{cases} H_0 : & \mu = \mu_0 \\ H_1 : & \mu \neq \mu_0 \end{cases}$$

Seguendo la falsariga di quanto visto sopra, procediamo come segue:

(a) Prendiamo $\overline{X} = \dfrac{\sum X_i}{n}$, che è stimatore della media del campione. Non conoscendo la distribuzione esatta delle variabili, consideriamo la distribuzione approssimata: sotto H_0 abbiamo che $\overline{X} \overset{a}{\sim} N(\mu_0, \sigma^2/n)$ (questo, come già detto varie volte, per il TLC).

(b) Scegliamo una regola di decisione: usando anche la distribuzione asintotica di $\overline{X}$ sotto H_0, individuiamo la *regione di rifiuto del test.* Dobbiamo trovare quindi due valori $h, k \in \mathbb{R}$ tali per cui *rifiuto* H_0 se $\overline{X_n} \leq h$ o $\overline{X_n} \geq k$.

(c) Scelto il livello di confidenza α, si ha che h, k vanno scelti in modo tale per cui

$$\alpha = P(\overline{X} \leq h \vee \overline{X} \geq k \mid \mu = \mu_0)$$

Grazie alla normalità della distribuzione asintotica, possiamo supporre che la distribuzione di $\overline{X_n}$ sia simmetrica, per lo meno da un certo n in poi (ricordiamo che, per campioni con numerosità superiori a 30, l'approssimazione è già buona).

Questo ci permette di scrivere

$$\frac{\alpha}{2} = P(\overline{X} \leq h \mid \mu = \mu_0) = P(\overline{X} \geq k \mid \mu = \mu_0)$$

Dato che

$$\frac{\overline{X} - \mu_0}{S_n/\sqrt{n}} \xrightarrow{D} N(0,1)$$

denotato con $z_{\alpha/2}$ il quantile di ordine $\alpha/2$ della normale standard, la regione critica (di rifiuto) sarà

$$C = \left\{ \underline{x} = (x_1, ..., x_n) \in X : \left| \frac{\overline{X} - \mu_0}{S_n/\sqrt{n}} \right| \geq z_{\alpha/2} \right\}$$

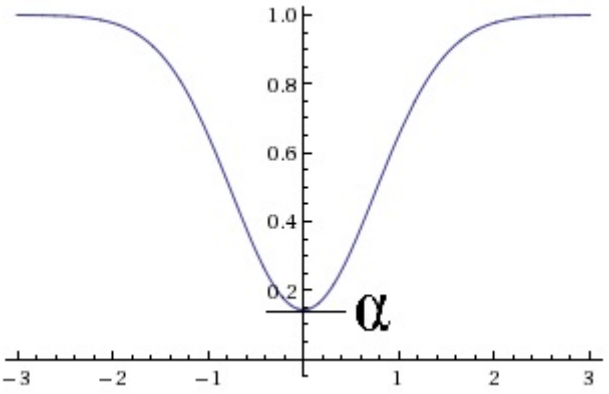
da cui

$$k = \frac{S_n}{\sqrt{n}} z_{\alpha/2} + \mu_0 (= -h)$$

Consideriamo inoltre la funzione di potenza associata, sempre sfruttando l'approssimazione normale ed il fatto che, per n grande, $S_n \approx \sigma$:

$$\begin{aligned} \gamma(\mu) &= P(\overline{X} \leq \mu_0 - \frac{S_n}{\sqrt{n}} z_{\alpha/2} \mid \mu \neq \mu_0) + P(\overline{X} \geq \mu_0 + \frac{S_n}{\sqrt{n}} z_{\alpha/2} \mid \mu \neq \mu_0) \\ &\approx \Phi\left(\frac{\sqrt{n}(\mu_0 - \mu)}{\sigma} - z_{\alpha/2} \right) + \left[1 - \Phi\left(\frac{\sqrt{n}(\mu_0 - \mu)}{\sigma} - z_{\alpha/2} \right) \right] \end{aligned}$$

Di seguito possiamo osservare il grafico della funzione:



Osservazione: i test costruiti usando t di Student sono più conservativi rispetto a quelli costruiti con approssimazione Normale, nel senso che è più facile che un test venga accettato usando la prima rispetto alla seconda. Questo è dato dal fatto che la distribuzione t di student ha le code più pesanti rispetto alla normale! (in particolare fissato un $\alpha \in (0,1)$, i quantili della normale standard $z_{\alpha/2}$ sono più piccoli dei quantili della t di Student con $n-1$ gradi di libertà $t_{\alpha/2;n-1}$, cioè $|z_{\alpha/2}| < |t_{\alpha/2;n-1}|$)

Lezione del 12/04, ultima modifica 20/05, Andrea Gadotti

Esempio di test t-Student per due campioni
Campione $(X_1, ..., X_{n_1})$ da $N(\mu_1, \sigma^2)$ e $(Y_1, ..., Y_{n_2})$ da $N(\mu_2, \sigma^2)$ indipendenti. La varianza σ è la stessa ma non è nota.
Supponiamo di avere elementi per pensare che:

$$\begin{cases} H_0: & \mu_1 = \mu_2 \\ H_1: & \mu_1 > \mu_2 \end{cases}$$

Abbiamo che $\overline{X}_1 - \overline{X}_2 \sim N(\mu_1 - \mu_2, \sigma_1/n_1 + \sigma_2/n_2)$.
Prendiamo ora

$$S_p^2 := \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n-2}$$

dove $n = n_1 + n_2$.
Abbiamo che:

$$T := \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{S_p\sqrt{1/n_1 + 1/n_2}} \overset{H_0}{\sim} t_{n-2}$$

dove abbiamo usato anche il fatto che S_p^2 è una chi-quadro con $n-2$ gradi di libertà (vedi pag. 19).
In conclusione, avremo che la nostra regione critica sarà $C = \{(\underline{x}, \underline{y}) \mid T \geq t_{n-2;\alpha}\}$.

Esempio con bernoulliana Abbiamo $(X_1, ..., X_n)$ campione casuale da $b(1,p)$ (ricordiamo: media p, varianza $p(1-p)$). Le nostre ipotesi sono:

$$\begin{cases} H_0: & p = p_0 \\ H_1: & p = p_1 \end{cases}$$

con $p_1 < p_0$. Consideriamo

$$\hat{p}_n := \frac{\sum X_i}{n} \overset{a}{\sim} N\left(p, \frac{p(1-p)}{n}\right)$$

Abbiamo quindi:

$$Z := \frac{\hat{p}_n - p_0}{\sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n}}}$$

A questo punto imponiamo: $\alpha = P(Z \leq z_\alpha \mid H_0)$.

Esempio di test sulla varianza Campione casuale $(X_1, ..., X_n)$ da $N(\mu, \sigma^2)$, μ e σ^2 non noti. Le nostre ipotesi sono:

$$\begin{cases} H_0: & \sigma^2 = \sigma_0^2 \\ H_1: & \sigma^2 = \sigma_1^2 \end{cases}$$

con $\sigma_1^2 > \sigma_0^2$. Notiamo che $\frac{n-1}{\sigma_0^2} S_n^2 \overset{H_0}{\sim} \chi^2_{n-1} =: W$. (Nota: come in molti esempi precedenti, il fatto che sia "intelligente" tirare fuori queste osservazioni che portano all'analisi di distribuzioni conosciute è "calato dall'alto", almeno per ora).
Imponiamo:

$$\alpha = P(S_n^2 \geq k \mid H_0) = P\left(\frac{n-1}{\sigma^2} S_n^2 \geq \frac{n-1}{\sigma^2} k \mid \sigma^2 = \sigma_0^2\right) = P\left(\frac{n-1}{\sigma_0^2} S_n^2 \geq \frac{n-1}{\sigma_0^2} k\right) = P\left(W \geq \frac{n-1}{\sigma_0^2} k\right)$$

Quindi $\frac{n-1}{\sigma_0^2} k = w_{\alpha;n-1}$. [3]
In conclusione, rifiuto H_0 se $W \geq w_{\alpha;n-1}$, ovvero se $S_n^2 \geq k = \frac{\sigma_0^2 \, w_{\alpha,n-1}}{n-1}$.

Esempio In riferimento al caso generale appena trattato, supponiamo di avere $n = 25$, $\sigma_0^2 = 15$, $\sigma_1^2 = 20$, $s_n^2 = 17,4$ e $\alpha = 0,05$. Allora:

$$k = w_{0,05;(25-1)} = w_{0,05;24} = 36,415 > 27,84 = \frac{25-1}{15} 17,4 = \frac{n-1}{\sigma_0^2} s_n^2 = w$$

In conclusione, non rifiutiamo H_0.

Esempio con due campioni normali Campione $(X_1, ..., X_{n_1})$ da $N(\mu_1, \sigma_1^2)$ e $(Y_1, ..., Y_{n_2})$ da $N(\mu_2, \sigma_2^2)$ indipendenti. μ_i e σ_i^2 non noti. Le nostre ipotesi sono:

$$\begin{cases} H_0: & \sigma_1^2 = \sigma_2^2 \\ H_1: & \sigma_1^2 \neq \sigma_2^2 \end{cases}$$

Nota: l'ipotesi dell'uguaglianza delle due varianze prende il nome di omoschedasticità.
Sappiamo che $S_i^2 \overset{P}{\to} \sigma_i^2$. Inoltre

$$W := \frac{S_2^2/\sigma_2^2}{S_1^2/\sigma_1^2} = \frac{\frac{(n_2-1)S_2^2}{\sigma_2^2} \frac{1}{n_2-1}}{\frac{(n_1-1)S_1^2}{\sigma_1^2} \frac{1}{n_1-1}} \sim F_{(n_2-1),(n_1-1)}$$

(vedi pag. 20)
Sotto H_0 abbiamo chiaramente che $W := \frac{S_2^2/\sigma_2^2}{S_1^2/\sigma_1^2} = \frac{S_2^2}{S_1^2}$. La nostra regola di decisione consisterà nel rifiutare H_0 a favore di H_1 se $\frac{S_2^2}{S_1^2}$ è "lontano" da 1, ovvero se $W < k_1$ o $W > k_2$, dove k_1 e k_2 dipendono dalla distribuzione di $W = \frac{S_2^2}{S_1^2}$ e dal valore di α. Perciò, dividendo equamente in due parti la probabilità di errore, le due equazioni risultano:

$$\alpha/2 = P(W > k_2 \mid \sigma_1^2 = \sigma_2^2) \quad \text{e} \quad 1 - \alpha/2 = P(W < k_1 \mid \sigma_1^2 = \sigma_2^2)$$

[3] Ricordiamo che per la distribuzione chi-quadro, con $w_{\alpha;n-1}$ intendiamo il valore tale che $P(\chi^2_{n-1} \geq w_{\alpha;n-1})$

In conclusione,

$$C = \{(\underline{x_1}, \underline{x_2}) \ : \ \frac{S_2^2}{S_1^2} < w_{(n_2-1),(n_1-1);\alpha/2} \quad \text{o} \quad \frac{S_2^2}{S_1^2} > w_{(n_2-1),(n_1-1);1-\alpha/2}$$

In riferimento al caso generale appena trattato, supponiamo di avere $n_1 = 14$, $n_2 = 10$ $s_1^2 = 17,4$, $\sigma_1^2 = 20$, $s_2^2 = 37,9$ e $\alpha = 0,05$.
Come nel caso generale, diciamo $W := S_2^2/S_1^2 \sim F_{(n_2-1),(n_1-1)}$.
Abbiamo che $w_{(10-1),(14-1);0,025} = 3,31$ e $w_{(10-1),(14-1);0,975} = 1/w_{13,19;0,025} = 1/3,76 = 0,26$.
Poiché $s_2^2/s_1^2 = 37,9/17,4 = 2,178$, decidiamo di non rifiutare H_0.

Parte II

Seconda parte del corso

Lezioni dal 15/04 al 06/05 comprese. Autore: Marco Peruzzetto.
Questa parte comprende ed amplia le cose viste a lezione, estendendo alcune dimostrazioni e osservazioni. Quanto non fatto in classe verrà denotato da un (*).

Definizione: Sia $\vec{X} \coloneqq (X_1, \dots X_n)$ un vettore casuale da distribuzione $F(\vec{x}, \theta)$ per $\theta \in \Theta$ e sia $f_{X_i}(x_i, \theta)$ la corrispondente funzione densità di ciascuna X_i, $\forall 1 \leq i \leq n$. Indicheremo con $\vec{x} = (x_1, \dots, x_n)$ una qualsiasi possibile determinazione del vettore $\vec{X}$. Essa conterrà tutta l'informazione in merito a θ. Possiamo allora definire la *Funzione di Verosimiglianza* come la funzione:

$$L(\theta, \vec{x}) \coloneqq f_{\vec{X}}(\vec{x}, \theta) = f_{(X_1,\dots,X_n)}(x_i, \dots, x_n, \theta) \ , \ \theta \in \Theta,$$

che rappresenta quindi la funzione di densità dell'intero vettore in dipendenza del parametro θ. Nel caso in cui il vettore casuale sia un campione casuale, allora tutte le variabili casuali di cui esso è composto saranno i.i.d., ragion per cui la funzione di massima verosimiglianza assumerà la seguente tipica forma:

$$L(\theta, \vec{x}) = \prod_{i=1}^{n} f_{X_i}(x_i, \theta) \ \ , \ \theta \in \Theta.$$

Esiste anche la *Funzione di Log-Verosimiglianza* definita come $l(\theta, \vec{x}) \coloneqq \log\big(L(\theta, \vec{x})\big)$.

Esempio: Sia $(X_1, \dots, X_n) \sim Poisson(\theta)$. Allora

$$L(\theta, \vec{x}) = \prod_{i=1}^{n} \frac{e^{-\theta}\theta^{x_i}}{x_i!} \mathbb{1}_{\mathbb{N}}(x_i) = \frac{e^{-n\theta}\theta^{\sum_{i=1}^{n} x_1}}{\prod_{i=1}^{n} x_i!} \prod_{i=1}^{n} \mathbb{1}_{\mathbb{N}}(x_i)$$

da cui

$$l(\theta, \vec{x}) = \log(\theta) \sum_{i=1}^{n} x_i - n\theta - \sum_{i=1}^{n} \log(x_i!).$$

Osservazioni:

- La funzione di verosimiglianza dà un valore alla probabilità che $\vec{x}$ provenga da $F_{\vec{X}}(\vec{x}, \theta)$ per tutti i differenti valori di $\theta \in \Theta$.
- Nella funzione di verosimiglianza è stato volontariamente invertito il parametro θ con il parametro $\vec{x}$ rispetto, ad esempio, alla funzione densità. La ragione si basa sulla diversa interpretazione della stessa: a tutti gli effetti la funzione di verosimiglianza non è altro che la funzione densità del vettore casuale $\vec{X}$. Quindi essa può essere vista in due modi diversi: il primo interpreta la funzione L come una funzione di $\vec{x}$, e quindi del risultato, una volta fissato il valore del parametro (perciò L esattamente la densità), mentre il secondo la interpreta come una funzione del parametro θ, per un fissato valore del risultato $\vec{x}$. Proprio in quest'ultimo caso ha senso parlare di verosimiglianza: il valore assunto da L indica quanto verosimilmente il valore di un parametro (θ) sia corretto rispetto al risultato che si possiede ($\vec{x}$).

- (*) Data una determinazione $\vec{x}$ di $\vec{X}$, la funzione $L(\cdot, \vec{x})$, essendo la densità del vettore casuale, esprime la probabilità che $\vec{X}$ assuma proprio il valore $\vec{x}$. Ciò avviene in modo diretto se le variabili componenti il vettore sono discrete e tramite integrazione se continue. Ha senso allora chiedersi, data una determinazione $\vec{x}_0$ di $\vec{X}$, quale sia (se esiste) un possibile valore $\theta_0 \in \Theta$ capace di massimizzare il valore di $L(\theta_0, \vec{x}_0)$. Massimizzare tale valore significa infatti per quanto detto, andare a massimizzare la probabilità che $\vec{X}$ assuma il valore $\vec{x}_0$. Ciò avverrà direttamente se il vettore casuale è discreto, ma anche se è continuo, e ciò banalmente grazie alla monotonia dell'integrale, in quanto, se riusciamo a massimizzare la funzione con θ anche l'integrale (ovviamente integrando in $d\vec{x}$) sarà massimo (rivedere).

- L'importanza di cercare il valore del parametro che massimizzi L fissata la determinazione risiede nel fatto che spesso in statistica si ha a che fare con poche determinazioni e si parte dunque dall'evidente presupposto che se il campionamento effettuato ci ha fornito quelle specifiche determinazioni, esse debbano essere mediamente le più probabili. Tale presupposto viene in effetti denominato *Principio di "Rational Belief"*. La probabilità che dato quel campione casuale si ottengano quelle determinazioni la immagineremo quindi come la massima possibile. Cercheremo dunque un $\theta \in \Theta$ che soddisfi a ciò. È inevitabile che attraverso la verosimiglianza si possano ottenere degli stimatori del parametro.

- La funzione di log-verosimiglianza è stata introdotta pressoché per il semplice motivo di semplificare i calcoli quando si cerca di andare a massimizzare la funzione di verosimiglianza. Essa risulta dunque essere comoda, in quanto, essendo il logaritmo una funzione strettamente crescente, il massimizzante di $l(\cdot, \cdot)$ coinciderà con quello di $L(\cdot, \cdot)$.

Esempio (Problema dei Pesci): Dato un lago, lo scopo è cercare di stimare la grandezza N della popolazione dei pesci che vi vivono. Un modo può essere il seguente: si pescano esattamente N_1 pesci, i quali vengono in qualche modo marcati. In seguito, dopo aver permesso un mescolamento, si esegue un'ulteriore pesca, di n pesci. Si nota che fra questi ve ne sono n_1 marcati. Vogliamo capire quale sia il valore di N più plausibile. Nel nostro caso avremo un vettore casuale composto da una sola variabile, ovvero $\vec{X} = (X)$, la quale ha valori in $\mathbb{N}$ (ed è quindi discreta) e restituisce i possibili valori di n_1. La sua densità sarà allora fornita in modo diretto e coincide con la funzione di verosimiglianza in quanto vi è una singola variabile casuale nel vettore. L'insieme dei parametri sarà anch'esso $\mathbb{N}$. Chiaramente vogliamo stimare il più plausibile valore di $\theta = N$. Avremo dunque: $L(N) \coloneqq L(N, n_1) = \mathbb{P}[X = n_1] = \frac{\binom{N}{n_1}\binom{N-N_1}{n-n_1}}{\binom{N}{n}}$. Per effettuare un esempio concreto: con $N_1 = 300$ e $n = 80$, se la nostra determinazione ottenuta fosse $n_1 = 30$, allora il parametro che massimizza la probabilità sarebbe $N \sim 1200$. È quindi plausibile che nel lago viva una quantità di pesci che si aggira effettivamente intorno ai 1200 esemplari.

Definizione: Assumiamo che la funzione di verosimiglianza sia dervabile per il parametro θ. Allora la funzione $S(\theta) \coloneqq \frac{\partial}{\partial\theta}(\theta, \vec{x})$ viene detta *Score Function*. L'equazione $S(\theta) = 0$ è chiamata *Equazione di Stima*.

Osservazione: Osserviamo che poiché la funzione densità di una qualsiasi variabile ca-

suale è sempre positiva o nulla, in quanto prodotto, lo dovrà essere anche la parte di $L(\cdot,\cdot)$ che non dipende da θ. Ne segue che, se vogliamo massimizzare la funzione di verosimiglianza, possiamo direttamente limitarci a considerare solo i valori di $\theta \in \Theta$ che rendano $L(\cdot,\cdot)$ strettamente positiva per ciascuna determinazione $\vec{x}$ fissata o scelta. Dunque si può restringere senza perdere generalità l'insieme Θ in modo da avere valori che non permettono a $L(\cdot,\cdot)$ di annullarsi. $S(\theta)$ risulta quindi avere una buona definizione, in quanto non è necessario effettuare ulteriori ipotesi su $l(\cdot,\cdot)$, essendo:

$$S(\theta) = \frac{\partial}{\partial\theta} l(\theta,\vec{x}) = \frac{\partial}{\partial\theta} \log\big(L(\theta,\vec{x})\big) = \frac{1}{L(\theta,\vec{x})} \frac{\partial}{\partial\theta} L(\theta,\vec{x}).$$

Definizione: La funzione di verosimiglianza induce uno stimatore del parametro θ. Esso sarà chiamato *Stimatore di Massima Verosimiglianza* ed è così definito: $\hat{\theta}_n = \hat{\theta}_n(\vec{X}) \coloneqq \arg\big\{\max_{\theta\in\Theta}\{L(\theta,\vec{X})\}\big\} = \arg\big\{\max_{\theta\in\Theta}\{l(\theta,\vec{X})\}\big\}$.

Osservazioni:

- Da ora in poi l'argomento delle funzioni $L(\cdot,\cdot)$ e $l(\cdot,\cdot)$ verrà interpretato a seconda della convenienza e del senso sia come $(\theta,\vec{x})$, ovvero come determinazione, oppure come $(\theta,\vec{X})$, ovvero vettore casuale. Osserviamo che in quest'ultimo caso, le funzioni L e l diventano esse stesse automaticamente variabili casuali o stimatori, che dir si voglia. Ciò si ripercuote inevitabilmente sulle funzioni ad esse collegate, ad esempio su $S(\theta)$.

- Se $L(\cdot,\cdot)$ o $l(\cdot,\cdot)$ sono derivabili rispetto a θ, la funzione indipendente da θ che risolve $\forall\vec{x}$ l'equazione di stima $S(\theta) = 0$ fornisce effettivamente lo stimatore di massima verosimiglianza. Oppure, equivalentemente, potremmo dire che lo stimatore di massima verosimiglianza $\hat{\theta}_n(\vec{X})$ è quello che soddisfa l'equazione $S\big(\hat{\theta}_n(\vec{X})\big) = 0$.

- In generale non vi è garanzia che lo stimatore di massima verosimiglianza esista, oppure, se esiste, che esso sia unico. Tuttavia nel caso di famiglie di densità che rispettino certe ipotesi di regolarità (per esempio le famiglie esponenziali) tale problema non si pone.

- Anche assumendo che tale stimatore esista e sia unico, non è detto che sia sempre ottenibile analiticamente. Talvolta sarà necessario ricorrere a metodi numerici per la risoluzione dell'equazione di stima.

Esempi:

1. Riprendiamo l'esempio precedente ove avevamo il campione casuale con variabili distribuite come poissoniane di parametro θ. La funzione di log-verosimiglianza era data da
$$l(\theta,\vec{x}) = \log(\theta)\sum_{i=1}^{n} x_i - n\theta - \sum_{i=1}^{n}\log(x_i!)$$
sicché otteniamo subito che
$$S(\theta) = \frac{1}{\theta}\sum_{i=1}^{n} x_i - n$$

Si deduce allora immediatamente l'equazione di stima $\frac{1}{\theta}\sum_{i=1}^{n} x_i - n = 0 \Rightarrow \theta = \frac{1}{n}\sum_{i=1}^{n} x_i$ Ma allora

$$\hat{\theta}_n(\vec{X}) = \overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

che è la media campionaria.

2. Sia $\vec{X} \coloneqq (X_1, \ldots, X_n) \sim U[(0, \theta)]$. Allora

$$f_X(x, \theta) \coloneqq \frac{1}{\theta}\mathbb{1}_{[0,\theta]}(x)$$

Perciò

$$L(\theta, \vec{x}) = \frac{1}{\theta^n}\prod_{i=1}^{n}\mathbb{1}_{[0,\theta]}(x_i) = \frac{1}{\theta^n}\mathbb{1}_{[X_{(n)},+\infty]}(\theta) \Rightarrow \hat{\theta}_n(\vec{X}) = X_{(n)}.$$

3. Sia $\vec{X} \coloneqq (X_1, \ldots, X_n) \sim \exp(\beta)$, $\beta > 0$. Allora $f_X(x, \theta) \coloneqq \beta e^{-\beta x}\mathbb{1}_{\mathbb{R}_+}(x)$. Perciò

$$L(\theta, \vec{x}) = \left(\beta^n e^{-\beta\sum_{i=1}^{n} x_i}\right)\mathbb{1}_{\mathbb{R}_+^n}(\vec{x}) \Rightarrow l(\theta, \vec{x}) = n\log(\beta) - \beta\sum_{i=1}^{n} x_i$$

Otterremo allora l'equazione di stima $0 = S(\beta) = n\beta - \sum_{i=1}^{n} x_i$, da cui subito si deduce che anche in questo caso $\hat{\theta}_n = \overline{X}_n$.

4. *(Troncamento)* Sia sempre $\vec{X} \coloneqq (X_1, \ldots, X_n) \sim \exp(\beta)$. Naturalmente ciascuna variabile casuale ha come codominio i reali non negativi. Possiamo supporre di aver effettuato gli n rilevamenti dal campione casuale e di essere riusciti a individuarne esattamente m puntualmente (che senza perdita di generalità immagineremo essere i primi m), mentre dei restanti $n - m$ immaginiamo di aver rilevato solamente che il loro valore supera una certa soglia fissta $T > 0$. Il campione contiene quindi due tipi di informazione da coniugare nella funzione di massima verosimiglianza, che avrà stavolta una forma un po' diversa. La indicheremo con L'. Otteniamo:

$$\begin{aligned}
L'(\beta, \vec{x}) &= \prod_{i=1}^{m} f_{X_i}(x_i, \beta) \cdot \prod_{i=m+1}^{n} \mathbb{P}[X_i > T] \\
&= \prod_{i=1}^{m} f_{X_i}(x_i, \beta) \cdot \prod_{i=m+1}^{n} (1 - F_{X_i}(T, \beta)) \\
&= \prod_{i=1}^{m} \beta e^{-\beta x_i} \cdot \prod_{i=m+1}^{n} \int_T^{+\infty} \beta e^{-\beta x_i} dx_i \\
&= \beta^m e^{-\beta\sum_{i=1}^{m} x_i} \cdot e^{-\beta(n-m)T}
\end{aligned}$$

Da cui

$$l'(\beta, \vec{x}) \coloneqq \log\left(L'(\beta, \vec{x})\right) = m\log(\beta) - \beta\sum_{i=1}^{m} x_i - \beta(n-m)T.$$

Inoltre possiamo definire anche qui una score function nel modo naturale:

$$S'(\beta) \coloneqq \frac{\partial}{\partial\beta} l'(\beta, \vec{x})$$

da cui, uguagliando a 0 si può ottenere l'equazione di stima

$$\frac{m}{\beta} - \sum_{i=1}^{m} x_i - (n-m)T = 0$$

Si deduce così lo stimatore di massima verosimiglianza con troncamento a T, dato da

$$\hat{\beta}'_n(\vec{X}) = \frac{\sum_{i=1}^{m} X_i + (n-m)T}{m}.$$

1.6 Efficienza

Dato uno stimatore T_n di un campione casuale $\vec{X} \coloneqq (X_1, \ldots, X_n)$ possiamo partire dal concetto di errore quadratico medio $\mathrm{MSE}_\theta(T_n) = \mathrm{Var}_\theta(T_n) + B^2_\theta(T_n)$. Lo scopo sarà quello di cercare stimatori che minimizzino il più possibile tale valore. Il problema presenta alcune difficoltà: per fare un piccolo esempio, sia $\theta_0 \in \Theta$ e consideriamo il seguente stimatore banale $U_n(\vec{X}) \coloneqq \theta_0$. È ora evidente che se da una parte $\mathrm{MSE}_{\theta_0}(U_n) = 0$, sicché nessun altro stimatore può essere uniformemente migliore di U_n, dall'altra appare chiaro che di un siffatto stimatore non ci si possa attendere molto, e nemmeno fidare, in quanto esso ignora completamente tutta l'informazione contenuta nel vettore casuale. La difficoltà di trovare stimatori che abbiano errore quadratico medio minimo è dunque legata a due aspetti principali: spesso la struttura di MSE è complicata in quanto contiene aspetti legati al parametro θ; inoltre la classe degli stimatori competitori di θ è quasi sempre troppo ampia. Cercheremo allora di semplificare il problema restringendo un po' il campo: considereremo solo gli stimatori non distorti, per andare poi a cercare tra questi quelli con varianza minima.

Esempio: Sia $\vec{X} \coloneqq (X_1, \ldots, X_n) \sim Poisson(\lambda)$. In tal caso si verifica subito che $\mathbb{E}[X] = \mathrm{Var}[X] = \lambda$. Ne segue che sia lo stimatore media campionaria $\overline{X}_n$ sia lo stimatore varianza campionaria S^2_n sono due stimatori non distorti di λ. Si ha tuttavia che $\mathrm{Var}[\overline{X}_n] = \frac{\lambda}{n} \leq \frac{\lambda}{n}\left(1 + \frac{2n\lambda}{n-1}\right) = \mathrm{Var}[S^2_n]$. Preferiremo dunque la media campionaria. Ma consideriamo ora il seguente stimatore così definito, per $a \in [0,1]$ fissato, $W_{n,a}(\vec{X}) \coloneqq a\overline{X}_n + (1-a)S^2_n$. Anch'esso è non distorto. Sorgono così due difficoltà da affrontare: ammesso che $\overline{X}_n$ sia migliore (i.e. con varianza più piccola) di S^2_n, esso è anche migliore di ogni stimatore $W_{n,a} \forall a$ oppure esso è il migliore tra tutti gli stimatori non distorti di λ? Esiste un limite inferiore alla varianza? Se infatti esso esistesse, darebbe operatività alla scelta dello stimatore, in quanto se trovassimo uno stimatore che raggiunge tale limite, sapremo che non è necessario cercare ulteriormente per migliorare le nostre possibilità. Ebbene, tale limite esiste sicuramente, sotto alcune ulteriori ipotesi di regolarità

da addure alla non distorsione per gli stimatori.

1.6.1 Teorema di Rao-Cramér

Definizione: Una *Famiglia Regolare* è una famiglia di densità che soddisfa le seguenti condizioni di regolarità:

1. *Condizione di Indentificabilità:* i valori delle densità sono distinti al variare del parametro, ovvero $\theta \neq \theta' \Longrightarrow f_X(x,\theta) \neq f_X(x,\theta')$.
2. Le funzioni densità hanno supporto comune $\forall \theta \in \Theta$ e il loro supporto non dipende in alcun modo dal parametro θ.
3. Le funzioni sono di classe C^2 rispetto alla variabile θ
4. Rispetto a θ, è lecito lo scambio tra le derivate e l'integrale.

Definizione: Sia $\vec{X} = (X_1, \ldots, X_n)$ un campione casuale. Allora la funzione

$$I : \Theta \longrightarrow \mathbb{R}$$
$$I(\theta) \coloneqq \mathbb{E}_\theta[S(\theta)^2] = \mathbb{E}_\theta\left[\left(\frac{\partial}{\partial\theta} l(\theta, \vec{X})\right)^2\right]$$

viene denominata *Informazione di Fisher* del campione casuale.

Osservazioni:

- Il prossimo teorema ci garantirà nel caso di famiglie regolari che il limite inferiore della varianza di un qualsiasi stimatore non distorto di θ è la quantità $\frac{1}{I(\theta)}$. Notiamo inoltre che più la varianza di uno stimatore si avvicina a tale quantità, più è significativa la sintesi dell'informazione circa θ contenuta nel vettore $\vec{X}$ realizzata dallo stimatore non distorto.

- Spesso si usano anche le seguenti notazioni per l'informazione di Fisher, ovvero $I(\theta)$, $I_n(\theta)$, $nI_1(\theta)$. Infatti dato un vettore casuale qualsiasi $\vec{X} = (X_1 \ldots, X_n)$, la sua funzione densità, ovvero $L(\cdot,\cdot)$ non si spezza necessariamente nel prodotto delle densità di ciascuna componente X_i. Ciò avviene invece nel caso in cui tutte le variabili casuali siano indipendenti: in tal caso si può scrivere $I_{\vec{X}}(\theta) = \sum_{i=1}^n I_{X_i}(\theta)$. Se poi siamo di fronte ad un campione casuale, allora le variabili casuali sono addirittura *i.i.d.*, e di conseguenza $I_{\vec{X}}(\theta) = nI_{X_1}(\theta)$, da cui la notazione.

- Si può dimostrare facilmente che, dato un campione casuale $\vec{X} \coloneqq (X_1, \ldots, X_n)$ con densità nella famiglia regolare, vale la seguente uguaglianza:

$$\mathbb{E}_\theta\left[\left(\frac{\partial}{\partial\theta} l(\theta, \vec{X})\right)^2\right] = -\mathbb{E}_\theta\left[\frac{\partial^2}{\partial\theta^2} l(\theta, \vec{X})\right]$$

In effetti, come già visto, si ha: $\frac{\partial}{\partial\theta}L(\theta,\vec{X}) = L(\theta,\vec{X})\frac{\partial}{\partial\theta}l(\theta,\vec{X})$. Perciò, derivando si ottiene subito che:

$$\begin{aligned}\frac{\partial^2}{\partial\theta^2}L(\theta,\vec{X}) &= L(\theta,\vec{X})\frac{\partial^2}{\partial\theta^2}l(\theta,\vec{X}) + \frac{\partial}{\partial\theta}L(\theta,\vec{X})\frac{\partial}{\partial\theta}l(\theta,\vec{X}) \\ &= L(\theta,\vec{X})\frac{\partial^2}{\partial\theta^2}l(\theta,\vec{X}) + L(\theta,\vec{X})\left(\frac{\partial}{\partial\theta}l(\theta,\vec{X})\right)^2.\end{aligned}$$

Si può quindi ricavare $\left(\frac{\partial}{\partial\theta}l(\theta,\vec{X})\right)^2 = \frac{1}{L(\theta,\vec{X})}\frac{\partial^2}{\partial\theta^2}L(\theta,\vec{X}) - \frac{\partial^2}{\partial\theta^2}l(\theta,\vec{X})$. Per provare l'asserto basterà dunque verificare che il valore di aspettazione del primo addendo del secondo termine dell'uguaglianza sia nullo. Si ha:

$$\begin{aligned}\mathbb{E}\Big[\frac{1}{L(\theta,\vec{X})}\frac{\partial^2}{\partial\theta^2}L(\theta,\vec{X})\Big] &= \int_{\mathbb{R}^n}\frac{1}{L(\theta,\vec{x})}\frac{\partial^2}{\partial\theta^2}L(\theta,\vec{x})\cdot L(\theta,\vec{x})\cdot d\vec{x} \\ &= \int_{\mathbb{R}^n}\frac{\partial^2}{\partial\theta^2}L(\theta,\vec{x})\cdot d\vec{x} \\ &= \frac{\partial^2}{\partial\theta^2}\int_{\mathbb{R}^n}L(\theta,\vec{x})\cdot d\vec{x} \\ &= \frac{\partial^2}{\partial\theta^2}1 = 0.\end{aligned}$$

Definizione: Sia $\vec{X} \coloneqq (X_1,\dots,X_n)$ un campione casuale e $T_n = T_n(\vec{X})$, $V_n = V_n(\vec{X})$ due stimatori non distorti di θ. Allora:

- Diremo *Efficienza assoluta o di Bahadur* di T_n il valore $\mathrm{eff}(T_n) \coloneqq \frac{\frac{1}{I(\theta)}}{\mathrm{Var}_\theta[T_n]}$.
- Diremo *Efficienza relativa* di T_n e V_n il valore $\mathrm{eff}(T_n,V_n) \coloneqq \frac{\mathrm{Var}_\theta[T_n]}{\mathrm{Var}_\theta[V_n]}$.
- Diremo che T_n è *Efficiente* se $\mathrm{eff}(T_n) = 1$. Nel caso in cui $\mathrm{eff}(T_n) > 1$ lo stimatore T_n si dirà anche *Super-Efficiente*. In generale, si dirà che T_n è più (meno) efficiente di V_n se $\mathrm{eff}(T_n,V_n) < (>)1$.
- Diremo che T_n è *Asintoticamente Efficiente* se $\lim_{n\to\infty}\mathrm{eff}(T_n) = 1$.

Teorema 15 (di Rao-Cramér)**.** *Sia $\vec{X} \coloneqq (X_1,\dots,X_n)$ un campione casuale di densità $f_{\vec{X}}(\vec{x},\theta)$ appartenente alla famiglia regolare con $\theta\in\Theta\subset\mathbb{R}$ un insieme di parametri. Sia poi $g:\Theta\longrightarrow\Theta$ una funzione derivabile e assumiamo l'informazione di Fisher $I(\theta)\neq 0$ $\forall\theta\in\Theta$. Allora, per qualsiasi stimatore $T_n = T_n(\vec{X})$ non distorto del parametro $g(\theta)$, vale* $\mathrm{Var}_\theta[T_n] \geq \big(g'(\theta)\big)^2\cdot\frac{1}{I(\theta)}$.

Dimostrazione. Poiché T_n è uno stimatore non distorto di $g(\theta)$, abbiamo:
$g(\theta) = \mathbb{E}_\theta[T_n] = \int_{\mathbb{R}^n}T_n(\vec{x})f_{\vec{X}}(\vec{x},\theta)d\vec{x} = \int_{\mathbb{R}^n}T_n(\vec{x})L(\theta,\vec{x})d\vec{x}$, con $\theta\in\Theta$. Perciò, derivando sotto il parametro θ e grazie alle ipotesi di regolarità otteniamo:

$$\begin{aligned}g'(\theta) &= \frac{\partial}{\partial\theta}\int_{\mathbb{R}^n}T_n(\vec{x})L(\theta,\vec{x})d\vec{x} = \int_{\mathbb{R}^n}T_n(\vec{x})\left(\frac{\partial}{\partial\theta}L(\theta,\vec{x})\right)d\vec{x} \\ &= \int_{\mathbb{R}^n}T_n(\vec{x})L(\theta,\vec{x})\left(\frac{\partial}{\partial\theta}l(\theta,\vec{x})\right)d\vec{x} = \int_{\mathbb{R}^n}T_n(\vec{x})\left(\frac{\partial}{\partial\theta}l(\theta,\vec{x})\right)f_{\vec{X}}(\vec{x},\theta)d\vec{x} \\ &= \mathbb{E}_\theta\Big[T_n(\vec{X})\cdot\frac{\partial}{\partial\theta}l(\theta,\vec{X})\Big].\end{aligned}$$

Osserviamo ora per prima cosa che:

$$\begin{aligned}\mathbb{E}_\theta\Big[\frac{\partial}{\partial\theta}l(\theta,\vec{X})\Big] &= \int_{\mathbb{R}^n}\left(\frac{\partial}{\partial\theta}l(\theta,\vec{x})\right)L(\theta,\vec{x})d\vec{x} = \int_{\mathbb{R}^n}\left(\frac{\partial}{\partial\theta}L(\theta,\vec{x})\right)d\vec{x}\\ &= \frac{\partial}{\partial\theta}\int_{\mathbb{R}^n}L(\theta,\vec{x})d\vec{x} = \frac{\partial}{\partial\theta}\int_{\mathbb{R}^n}f_{\vec{X}}(\vec{x},\theta)d\vec{x}\\ &= \frac{\partial}{\partial\theta}1 = 0\end{aligned}$$

Ne seguono direttamente le due seguenti relazioni:

- $\mathrm{Cov}_\theta\left[T_n(\vec{X}),\frac{\partial}{\partial\theta}l(\theta,\vec{X})\right] = \mathbb{E}_\theta\big[T_n(\vec{X})\cdot\frac{\partial}{\partial\theta}l(\theta,\vec{X})\big] - \mathbb{E}_\theta\big[T_n(\vec{X})\big]\cdot\mathbb{E}_\theta\big[\frac{\partial}{\partial\theta}l(\theta,\vec{X})\big] = \mathbb{E}_\theta\big[T_n(\vec{X})\cdot\frac{\partial}{\partial\theta}l(\theta,\vec{X})\big] - \mathbb{E}_\theta\big[T_n(\vec{X})\big]\cdot 0 = \mathbb{E}_\theta\big[T_n(\vec{X})\cdot\frac{\partial}{\partial\theta}l(\theta,\vec{X})\big] = g'(\theta)$;
- $\mathrm{Var}_\theta\left[\frac{\partial}{\partial\theta}l(\theta,\vec{X})\right] = \mathbb{E}_\theta\left[\left(\frac{\partial}{\partial\theta}l(\theta,\vec{X})\right)^2\right] - \mathbb{E}_\theta\big[\frac{\partial}{\partial\theta}l(\theta,\vec{X})\big]^2 = \mathbb{E}_\theta\left[\left(\frac{\partial}{\partial\theta}l(\theta,\vec{X})\right)^2\right]$, di conseguenza $\mathrm{Var}_\theta\left[\frac{\partial}{\partial\theta}l(\theta,\vec{X})\right] = I_n(\theta)$.

D'altra parte, dalla disuguaglianza di Cauchy-Schwarz abbiamo:

$$\begin{aligned}\big(g'(\theta)\big)^2 &= \mathrm{Cov}_\theta\left[T_n(\vec{X}),\frac{\partial}{\partial\theta}l(\theta,\vec{X})\right]^2\\ &\le \mathrm{Var}_\theta[T_n(\vec{X})]\cdot\mathrm{Var}_\theta\left[\frac{\partial}{\partial\theta}l(\theta,\vec{X})\right] = \mathrm{Var}_\theta[T_n(\vec{X})]\cdot I_n(\theta),\end{aligned}$$

grazie alle relazioni appena introdotte. La tesi segue subito, ricordando che sia la varianza che l'informazione di Fisher sono quantità positive. □

Controesempio: Le ipotesi di regolarità del teorema sono necessarie. Consideriamo infatti il cmpione casuale $\vec{X} \coloneqq (X_1,\dots,X_n) \sim U([0,\theta])$. La sua densità non appartiene alla famiglia regolare in quanto ha il supporto dipendente dal parametro θ. Uno stimatore non distorto di θ abbiamo già visto essere $T_n(\vec{X}) \coloneqq \frac{n-1}{n}X_{(n)}$. Tuttavia $\mathrm{Var}[T_n] < \frac{1}{I(\theta)}$ e di conseguenza è stimatore super-efficiente. La tesi del teorema non è dunque valida in questo caso.

Lemma 1. *Sotto le usuali condizioni di regolarità, esiste uno stimatore non distorto T_n di θ efficiente, ossia tale che la sua varianza raggiunge il limite inferiore di Rao-Cramér, se e solo se $S(\theta) = \frac{\partial}{\partial\theta}l(\theta,\vec{X}) = I_n(\theta)\Big(T_n(\vec{X}) - \theta\Big)$.*

Dimostrazione. Grazie alla disuguaglianza di Cauchy-Schwarz abbiamo la seguente relazione $\mathrm{Cov}^2_\theta[T_n(\vec{X}),\frac{\partial}{\partial\theta}l(\theta,\vec{X})] \le \mathrm{Var}_\theta[T_n(\vec{X})]\cdot\mathrm{Var}_\theta[\frac{\partial}{\partial\theta}l(\theta,\vec{X})]$, nella quale sussiste l'uguaglianza sse vi è linarità tra i due termini, ovvero sse $\exists a,b\in\mathbb{R}$ tali che $\frac{\partial}{\partial\theta}l(\theta,\vec{X}) = a + bT_n(\vec{X})$. Come già calcolato nella precedente dimostrazione, il valore di aspettazione del primo membro dell'uguaglianza è nullo, perciò $0 = \mathbb{E}_\theta[\frac{\partial}{\partial\theta}l(\theta,\vec{X})] = \mathbb{E}_\theta[a + bT_n(\vec{X})] = \mathbb{E}_\theta[a] + \mathbb{E}_\theta[bT_n(\vec{X})] = a + b\theta \Rightarrow a = -b\theta$. Quindi $\frac{\partial}{\partial\theta}l(\theta,\vec{X}) = b(T_n(\vec{X}) - \theta)$. Se moltiplichiamo tutto per $\frac{\partial}{\partial\theta}l(\theta,\vec{X})$ abbiamo che $\left(\frac{\partial}{\partial\theta}l(\theta,\vec{X})\right)^2 = bT_n(\vec{X})\frac{\partial}{\partial\theta}l(\theta,\vec{X}) - b\theta\frac{\partial}{\partial\theta}l(\theta,\vec{X})$. Calcolando infine nuovamente il valore di aspettazione e riprendendo alcuni risultati ottenuti dalla dimostrazione del teorema di Rao-Cramér abbiamo che:

$I(\theta) = \mathbb{E}_\theta\Big[\Big(\frac{\partial}{\partial\theta}l(\theta,\vec{X})\Big)^2\Big] = b\mathbb{E}_\theta[T_n(\vec{X})\frac{\partial}{\partial\theta}l(\theta,\vec{X})] - b\theta\mathbb{E}_\theta[\frac{\partial}{\partial\theta}l(\theta,\vec{X})] = b\cdot 1 - b\theta\cdot 0$ e di conseguenza si ha $b = I(\theta)$, da cui si deduce immediatamente la tesi.

□

Esempio: Consideriamo ancora $\vec{X} \coloneqq (X_1,\dots,X_n) \sim Poisson(\lambda)$. Sappiamo che la sua densità appartiene alla famiglia regolare. Avevamo introdotto $\forall a \in [0,1]$ fissato gli stimatori non distorti $W_{n,a}(\vec{X}) \coloneqq a\overline{X}_n + (1-a)S_n^2$ e ci eravamo chiesti quale fosse il migliore. Ebbene, tra tutti essi, la risposta è proprio $W_{n,1} = \overline{X}_n$, la media campionaria. Infatti si ha, come già visto, che la score function è data da $S(\lambda) = -n + \frac{1}{\lambda}\sum_{i=1}^n X_i = n\left(\frac{1}{\lambda}\overline{X}_n - 1\right)$. Se ora calcoliamo $I_n(\lambda) = -\mathbb{E}_\lambda[\frac{\partial^2}{\partial\lambda^2}l(\lambda,\vec{X})] = -\mathbb{E}_\lambda[\frac{d}{d\lambda}S(\lambda)] = -\mathbb{E}_\lambda[-\frac{1}{\lambda^2}\overline{X}_n] = \frac{1}{\lambda^2}\cdot n\lambda = \frac{n}{\lambda}$, otteniamo che $S(\lambda) = (\overline{X}_n - \lambda)\cdot\frac{n}{\lambda} = (\overline{X}_n - \lambda)\cdot I(\lambda)$ e possiamo concludere grazie il Lemma 1.

Lemma 2. *Sotto le usuali ipotesi di regolarità, sia $I(\theta) \neq 0\ \forall\theta\in\Theta$ e supponiamo che esista uno stimatore T_n non distorto di θ efficiente. Se $\hat{\theta}_n$ è lo stimatore di massima verosimiglianza di θ, allora vale $T_n = \hat{\theta}_n$.*

Dimostrazione. Il limite inferiore di Rao-Cramér non è una quantità nulla. Inoltre come già osservato e grazie al Lemma 1 si ha:

$$0 = S\big(\hat{\theta}_n(\vec{X})\big) = \big(T_n(\vec{X}) - \hat{\theta}_n(\vec{X})\big)I_n\big(\hat{\theta}_n(\vec{X})\big),$$

da cui $T_n(\vec{X}) - \hat{\theta}_n(\vec{X}) = 0$ e quindi la tesi.

□

Controesempio: Non sempre lo stimatore di massima verosimiglianza è anche stimatore efficiente, e dunque, per il Lemma 2, non sempre esiste uno stimatore efficiente. Sia infatti $\vec{X} \coloneqq (X_1,\dots,X_n) \sim f_X(x,\theta) \coloneqq \theta x^{\theta-1}\cdot\mathbb{1}_{(0,1)}(x)$ campione casuale, con $\theta > 0$. Ora, $\frac{\partial^2}{\partial\theta^2}\log\big(f(x,\theta)\big) = -\frac{1}{\theta^2} \Rightarrow I_1(\theta) = -\mathbb{E}[-\frac{1}{\theta^2}] = \frac{1}{\theta^2} \Rightarrow I_n(\theta) = \frac{n}{\theta^2}$. Però $S(\theta) = \frac{\partial}{\partial\theta}l(\theta,\vec{X}) = \frac{\partial}{\partial\theta}\log\Big(\prod_{i=1}^n \theta X_i^{\theta-1}\Big) = \frac{n}{\theta} + \sum_{i=1}^n\log(X_i)$. L'equazione di stima $S(\theta) = 0$ ci fornisce allora $\hat{\theta}_n(\vec{X}) = \frac{n}{\sum_{i=1}^n\log(X_i)}$, lo stimatore di massima verosimiglianza. Vogliamo ora trovare la sua distribuzione. Definiamo innanzi tutto il nuovo vettore casuale $\vec{Y} \coloneqq (Y_1,\dots,Y_n)$ dove $\forall i = 1..n$ si ha $Y_i \coloneqq \log(X_i)$. Osserviamo che il logaritmo è una funzione monotona crescente, e possiamo applicare il teorema 1.1 per ottenere che la densità delle nuove variabili è $f_Y(y,\theta) = \theta(e^{-y})^{\theta-1}|-e^{-y}|\cdot\mathbb{1}_{\mathbb{R}_+}(y) = \theta e^{-\theta y}\mathbb{1}_{\mathbb{R}_+}(y)$, e $\theta > 0$. Dunque, $\vec{Y} \sim G(\alpha = 1, \beta = \frac{1}{\theta})$. Poiché $\vec{X}$ è un vettore indipendente, segue necessariamente che anche $\vec{Y}$ lo sia; quindi, grazie alla proprietà di riproducibilità della densità Gamma $W \coloneqq \sum_{i=1}^n Y_i \sim G(\alpha' = n, \beta = \frac{1}{\theta})$. Si può mostrare che:

$$\mathbb{E}[W^k] = \frac{(n+k-1)!}{\theta(n-1)!}.$$

Ricordando che $\hat{\theta}_n = nW^{-1}$ possiamo calcolare subito i valori di aspettazione

- $\mathbb{E}_\theta[\hat{\theta}_n] = \mathbb{E}_\theta[nW^{-1}] = n\mathbb{E}_\theta[W^{-1}] = \frac{n}{n-1}\theta \neq \theta$, perciò è stimatore distorto, anche se asintoticamente non distorto.
- $\mathbb{E}[(\hat{\theta}_n)^2] = \mathbb{E}[n^2W^{-2}] = n^2\mathbb{E}[W^{-2}] = \frac{\theta^2n^2}{(n-2)(n-1)}$

e dunque $\mathrm{Var}[\hat{\theta}_n] = \mathbb{E}[(\hat{\theta}_n)^2] - \mathbb{E}[\hat{\theta}_n]^2 = \frac{n^2\theta^2}{(n-1)^2(n-2)^2} > \frac{1}{I(\theta)} = \frac{\theta}{n}$. Ne segue che $\hat{\theta}_n$ non è stimatore efficiente di θ anche se $\mathrm{eff}(\hat{\theta}_n) \xrightarrow[n\to\infty]{} 1$.

1.6.2 Estensione a un vettore di parametri:

Possiamo, al posto di un singolo parametro, andare a considerare un vettore di parametri $\vec{\theta} \coloneqq (\theta_1, \ldots, \theta_k) \in \Theta^k$, $\Theta \subset \mathbb{R}$ che indicizza la distribuzione di una variabile casuale X. Ad esempio la distribuzione Gamma dipende da due parametri solitamente indicati con α e β. In particolare, modellare un fenomeno con un numero di parametri che sia il più piccolo possibile assume un valore importante per quanto riguarda la stabilità degli stimatori. A ciò è stato dato il nome piuttosto eloquente di *Principio di Parsimonia*. Nel caso di un vettore di parametri si potrà allora estendere il concetto di Informazione di Fisher ottenendo una matrice.

Definizione. Sia $\vec{X}$ un campione casuale e $\vec{\theta} \coloneqq (\theta_1, \ldots, \theta_k)$ un campione di parametri. Allora la *Matrice di Informazione di Fisher* è la matrice $I(\vec{\theta}) \in \mathcal{M}(k \times k, \mathbb{R})$ il cui i-j-esimo elemento è dato dal numero

$$\mathbb{E}_{\vec{\theta}}\Big[\frac{\partial}{\partial\theta_i} l(\vec{\theta}, \vec{X}) \cdot \frac{\partial}{\partial\theta_j} l(\vec{\theta}, \vec{X})\Big].$$

Proposizione. (*) *Per ogni vettore casuale $\vec{X}$ si ha che la matrice di informazione di Fisher è simmetrica e semidefinita positiva.*

Dimostrazione. (*) Il fatto che sia simmetrica è pressoché immediato e viene direttamente dalla definizione. Per mostrare che è anche semidefinita positiva, sia $w(\vec{\theta}) \coloneqq \big(\frac{\partial}{\partial\theta_1} l(\vec{\theta}, \vec{X}), \ldots, \frac{\partial}{\partial\theta_k} l(\vec{\theta}, \vec{X})\big)$. Si vede subito che $I(\vec{\theta}) = \mathbb{E}_{\vec{\theta}}[w(\vec{\theta})^t \cdot w(\vec{\theta})]$. Sia ora $\vec{u} \in \mathbb{R}^k \setminus \{0\}$. Dobbiamo mostrare che $(u \cdot I(\vec{\theta}) \cdot u^t) \geq 0$. Ebbene, sfruttando la linearità del valore di aspettazione, otteniamo che:

$$\begin{aligned} uI(\vec{\theta})u^t &= u\mathbb{E}_{\vec{\theta}}[w(\vec{\theta})^t \cdot w(\vec{\theta})]u^t = \mathbb{E}_{\vec{\theta}}[u \cdot w(\vec{\theta})^t \cdot w(\vec{\theta}) \cdot u^t] \\ &= \mathbb{E}_{\vec{\theta}}[u \cdot w(\vec{\theta})^t \cdot ((u \cdot w(\vec{\theta})^t)^t] \\ &= \mathbb{E}_{\vec{\theta}}[\|u \cdot w(\vec{\theta})^t\|^2] \geq 0. \end{aligned}$$

□

Osservazione: Osserviamo che le ipotesi di regolarità definite per il caso unidimensionale, possono essere espanse al caso k-dimensionale nel modo più naturale, ovvero supponendo che esse valgano per ciascuno dei parametri θ_i, $\forall 1 \leq i \leq k$. Ebbene, sotto le usuali ipotesi di regolarità, si ottiene facilmente con quanto già mostrato che $I(\theta)_{ij} = -\mathbb{E}_{\vec{\theta}}\big[\frac{\partial^2}{\partial\theta_i\partial\theta_j} l(\vec{\theta}, \vec{x})\big]$. Osserviamo che vi è coerenza con la simmetria della matrice di informazione: essendo le densità di classe C^2, vale il teorema di Schwarz sullo scambio delle derivate.

Lemma. (*) *Siano $A \in \mathcal{M}(n \times n, \mathbb{R})$ una matrice simmetrica e definita positiva, e $b \in \mathbb{R}^n$.*

Allora, se definiamo la funzione

$$\begin{array}{rcl} f : \mathbb{R}^n & \longrightarrow & \mathbb{R} \\ f(x) & \coloneqq & x \cdot A \cdot x^t - 2b \cdot x^t \end{array}$$

abbiamo che f ha un unico punto di minimo $\hat{x} \coloneqq b \cdot A^{-1}$.

Dimostrazione. (*) Scriviamo per semplicità $x = (x_1, \ldots, x_n)$, $b = (b_1, \ldots, b_n)$ e $A = (a_{ij})_{ij}$. utilizzeremo il metodo della matrice hessiana. Abbiamo che:

$$f(x) = \sum_{i,j=1}^{n} x_i a_{ij} x_j - 2\sum_{h=1}^{n} b_h x_h = \sum_{k=1}^{n} a_{kk} x_k^2 + 2\sum_{i=2}^{n}\sum_{j=1}^{i-1} x_i a_{ij} x_j - 2\sum_{h=1}^{n} b_h x_h$$

dove l'ultima uguaglianza viene direttamente dal fatto che A è una matrice simmetrica. Definendo ora $\sum_1^0 \coloneqq 0$, calcoliamo la r-esima derivata, $\forall 1 \leq r \leq n$:

$$\begin{aligned} \frac{\partial}{\partial x_r} f(x) &= 2a_{rr}x_r + 2\sum_{j=1}^{r-1} a_{rj}x_j + 2\sum_{i=r+1}^{n} x_i a_{ir} - 2b_r \\ &= 2a_{rr}x_r + 2\sum_{j=1}^{r-1} a_{rj}x_j + 2\sum_{i=r+1}^{n} x_i a_{ri} - 2b_r \\ &= 2a_{rr}x_r + 2\sum_{s=1, s\neq r}^{n} a_{sr}x_s - 2b_r \\ &= 2\sum_{s=1}^{n} a_{rs}x_s - 2b_r = 2(A_r \cdot x^t) - 2b_r, \end{aligned}$$

ove A_r è la r-esima riga della matrice A. Ora, per trovare i possibili punti stazionari sarà necessario eguagliare a 0 tutte le derivate parziali e risolvere il sistema. Non avendo termini quadratici esso sarà lineare:

$$\left\{ \begin{array}{l} \frac{\partial}{\partial x_1} f(x) = 0 \\ \vdots \\ \frac{\partial}{\partial x_n} f(x) = 0 \end{array} \right. \Leftrightarrow \left\{ \begin{array}{l} 2(A_1 \cdot x^t) - 2b_1 = 0 \\ \vdots \\ 2(A_n \cdot x^t) - 2b_n = 0 \end{array} \right. \Leftrightarrow \left\{ \begin{array}{l} A_1 \cdot x^t = b_1 \\ \vdots \\ A_n \cdot x^t = b_n \end{array} \right. \Leftrightarrow A \cdot x^t = b^t.$$

Allora, poiché la matrice A è invertibile, otterremo l'unico punto stazionario $\hat{x}^t = A^{-1} \cdot b^t$, ed essendo la matrice simmetrica, sarà $\hat{x} = b \cdot A^{-1}$. Per verificare adesso che si tratta di un punto di minimo, andiamo a calcolare la matrice hessiana di tutte le derivate seconde. Palesemente f è una funzione di classe C^∞, ragion per cui deve valere il teorema di Schwarz sullo scambio delle derivate seconde. In generale $\forall 1 \leq s, t \leq n$, si avrà:

$$\frac{\partial^2}{\partial x_s \partial x_t} f(x) = \frac{\partial}{\partial x_s}\left(\frac{\partial}{\partial x_t} f(x)\right) = \frac{\partial}{\partial x_s}\left(2\sum_{i=1}^{n} a_{ti}x_i - 2b_t\right) = 2a_{ts}.$$

Ne segue quindi che la matrice hessiana di f sarà $\forall x \in \mathbb{R}^n$ data da $2A$. Essa risulta quindi definita positiva e conferma di conseguenza che $\hat{x}$ è un punto di minimo. □

Teorema 16 (di Rao-Cramér)**.** *Sia $\vec{X} \coloneqq (X_1, \ldots, X_n)$ un campione casuale di densità $f_{\vec{X}}(\vec{x}, \vec{\theta})$ appartenente alla famiglia regolare e $\theta \in \Theta \subset \mathbb{R}^k$ un insieme di parametri. Sia poi $g : \Theta \longrightarrow \mathbb{R}$ una funzione derivabile e assumiamo che la matrice di informazione di Fisher $I(\theta)$ sia invertibile $\forall \theta \in \Theta$. Definiamo ora il vettore $\gamma(\vec{\theta}) \coloneqq \big(\frac{\partial}{\partial\theta_1} g(\vec{\theta}), \ldots, \frac{\partial}{\partial\theta_n} g(\vec{\theta})\big)$. Allora, per ogni stimatore $T = T(\vec{X})$ non distorto del numero $g(\vec{\theta})$ si ha che* $\mathrm{Var}[T] \geq \gamma(\vec{\theta}) \cdot I^{-1}(\vec{\theta}) \cdot \gamma(\vec{\theta})^t$.

Dimostrazione. (*) La dimostrazione si articola sfruttando alcuni passaggi già utilizzati nel caso unidimensionale. Innanzi tutto $\forall 1 \leq j \leq k$ si ha:

$$\begin{aligned}
\frac{\partial}{\partial\theta_j} g(\vec{\theta}) &= \frac{\partial}{\partial\theta_j} \mathbb{E}_{\vec{\theta}}[T] = \frac{\partial}{\partial\theta_j} \int_{\mathbb{R}^n} T(\vec{x}) f_{\vec{X}}(\vec{x}, \vec{\theta}) d\vec{x} = \frac{\partial}{\partial\theta_j} \int_{\mathbb{R}^n} T(\vec{x}) L(\vec{\theta}, \vec{x}) d\vec{x} \\
&= \int_{\mathbb{R}^n} T(\vec{x}) \frac{\partial}{\partial\theta_j} L(\vec{\theta}, \vec{x}) d\vec{x} = \int_{\mathbb{R}^n} T(\vec{x}) \left(\frac{\partial}{\partial\theta_j} l(\vec{\theta}, \vec{x}) \right) L(\vec{\theta}, \vec{x}) d\vec{x} \\
&= \mathbb{E}_{\vec{\theta}}\Big[T(\vec{x}) \cdot \left(\frac{\partial}{\partial\theta_j} l(\vec{\theta}, \vec{x}) \right) \Big].
\end{aligned}$$

Osserviamo allora che essendo

$$\begin{aligned}
\mathbb{E}_{\vec{\theta}}\Big[\frac{\partial}{\partial\theta_j} l(\vec{\theta}, \vec{X}) \Big] &= \int_{\mathbb{R}^n} \frac{\partial}{\partial\theta_j} l(\vec{\theta}, \vec{x}) L(\vec{\theta}, \vec{x}) d\vec{x} = \int_{\mathbb{R}^n} \left(\frac{\partial}{\partial\theta_j} L(\vec{\theta}, \vec{x}) \right) d\vec{x} \\
&= \frac{\partial}{\partial\theta_j} \int_{\mathbb{R}^n} L(\vec{\theta}, \vec{x}) d\vec{x} = \frac{\partial}{\partial\theta_j} \int_{\mathbb{R}^n} f_{\vec{X}}(\vec{x}, \vec{\theta}) d\vec{x} \\
&= \frac{\partial}{\partial\theta_j} 1 = 0,
\end{aligned}$$

$\mathrm{Cov}_{\vec{\theta}}\big[T_n(\vec{X}), \frac{\partial}{\partial\theta_j} l(\vec{\theta}, \vec{X})\big] = \mathbb{E}_{\vec{\theta}}\big[T_n(\vec{X}) \cdot \frac{\partial}{\partial\theta_j} l(\vec{\theta}, \vec{X})\big] - \mathbb{E}_{\vec{\theta}}\big[T_n(\vec{X})\big] \cdot \mathbb{E}_{\vec{\theta}}\big[\frac{\partial}{\partial\theta_j} l(\vec{\theta}, \vec{X})\big] = \mathbb{E}_{\vec{\theta}}\big[T_n(\vec{X}) \cdot \frac{\partial}{\partial\theta_j} l(\vec{\theta}, \vec{X})\big] - \mathbb{E}_{\vec{\theta}}\big[T_n(\vec{X})\big] \cdot 0 = \mathbb{E}_{\vec{\theta}}\big[T_n(\vec{X}) \cdot \frac{\partial}{\partial\theta_j} l(\vec{\theta}, \vec{X})\big] = \frac{\partial}{\partial\theta_j} g(\vec{\theta})$.

Fatto ciò, sia ora $\vec{c} = (c_1, \ldots, c_k) \in \mathbb{R}^k$ e definiamo ora la seguente funzione: $W(\vec{x}, \vec{\theta}) \coloneqq \sum_{i=1}^k c_i \frac{\partial}{\partial\theta_i} l(\vec{\theta}, \vec{x})$. È chiaro che $W(\vec{X}, \vec{\theta})$ sarà allora una variabile casuale. Vogliamo calcolare il valore di $\mathrm{Var}[T(\vec{X}) - W(\vec{X}, \vec{\theta})]$. Per prima cosa osserviamo che, sfruttando la linearità del valore atteso,

$$\begin{aligned}
\mathrm{Var}[T - W] &= \mathbb{E}[(T - W)^2] - \mathbb{E}[(T - W)]^2 \\
&= \mathbb{E}[T^2 - 2TW + W^2] - (\mathbb{E}[T] - \mathbb{E}[W])^2 \\
&= \mathbb{E}[T^2] - 2\mathbb{E}[TW] + \mathbb{E}[W^2] - \mathbb{E}[T]^2 + 2\mathbb{E}[T]\mathbb{E}[W] - \mathbb{E}[W]^2 \\
&= \mathrm{Var}[T] - 2\,\mathrm{Cov}[T, W] + \mathrm{Var}[W]
\end{aligned}$$

Adesso osserviamo invece che i conti si possono semplificare in quanto

$$\mathbb{E}[W(\vec{X}, \vec{\theta})] = \mathbb{E}\Big[\sum_{i=1}^k c_i \frac{\partial}{\partial\theta_i} l(\vec{\theta}, \vec{X}) \Big] = \sum_{i=1}^k c_i \mathbb{E}\Big[\frac{\partial}{\partial\theta_i} l(\vec{\theta}, \vec{X}) \Big] = \sum_{i=1}^k c_i \cdot 0 = 0.$$

Perciò otteniamo subito che:

$$\begin{aligned}
\mathrm{Cov}[T(\vec{X}), W(\vec{X},\vec{\theta})] &= \mathbb{E}[T(\vec{X})W(\vec{X},\vec{\theta})] - \mathbb{E}[T(\vec{X})]\mathbb{E}[W(\vec{X},\vec{\theta})] \\
&= \mathbb{E}[T(\vec{X})W(\vec{X},\vec{\theta})] \\
&= \mathbb{E}\Big[\sum_{i=1}^{k} c_i T(\vec{X})\frac{\partial}{\partial\theta_i} l(\vec{\theta},\vec{X})\Big] = \sum_{i=1}^{k} c_i \mathbb{E}\Big[T(\vec{X})\frac{\partial}{\partial\theta_i} l(\vec{\theta},\vec{X})\Big] \\
&= \sum_{i=1}^{k} c_i \,\mathrm{Cov}_{\vec{\theta}}\left[T_n(\vec{X}), \frac{\partial}{\partial\theta_i} l(\vec{\theta},\vec{X})\right] = \sum_{i=1}^{k} c_i \frac{\partial}{\partial\theta_i} g(\vec{\theta}) \\
&= \vec{c}\cdot\gamma(\vec{\theta})^t;
\end{aligned}$$

$$\begin{aligned}
\mathrm{Var}[W(\vec{X},\vec{\theta})] &= \mathbb{E}[\big(W(\vec{X},\vec{\theta})\big)^2] - \mathbb{E}[W(\vec{X},\vec{\theta})]^2 = \mathbb{E}[\big(W(\vec{X},\vec{\theta})\big)^2] \\
&= \mathbb{E}\Big[\big(\sum_{i=1}^{k} c_i \frac{\partial}{\partial\theta_i} l(\vec{\theta},\vec{X})\big)^2\Big] \\
&= \mathbb{E}\Big[\sum_{i=1}^{k}\sum_{j=1}^{k}\left(c_i\frac{\partial}{\partial\theta_i} l(\vec{\theta},\vec{X})\right)\left(\frac{\partial}{\partial\theta_j} l(\vec{\theta},\vec{X})c_j\right)\Big] \\
&= \sum_{i=1}^{k}\sum_{j=1}^{k} c_i \mathbb{E}\Big[\frac{\partial}{\partial\theta_i} l(\vec{\theta},\vec{X})\cdot\frac{\partial}{\partial\theta_j} l(\vec{\theta},\vec{X})\Big]c_j \\
&= \sum_{i=1}^{k}\sum_{j=1}^{k} c_i I_{ij}(\vec{\theta})c_j = \vec{c}\cdot I(\vec{\theta})\cdot(\vec{c})^t.
\end{aligned}$$

Osservando infine che la varianza di una qualsiasi variabile casuale è un numero sempre positivo, otteniamo la relazione definitiva:

$$0 \le \mathrm{Var}[T(\vec{X}) - W(\vec{X},\vec{\theta})] = \mathrm{Var}[T(\vec{X})] + \vec{c}\cdot I(\vec{\theta})\cdot(\vec{c})^t - 2\vec{c}\cdot\gamma(\vec{\theta})^t.$$

Poiché ciò deve valere $\forall\vec{c}\in\mathbb{R}^k$ possiamo allore scrivere:

$$0 \le \min_{\vec{c}\in\mathbb{R}^k}\{\mathrm{Var}[T(\vec{X}) - W(\vec{X},\vec{\theta})]\} = \mathrm{Var}[T(\vec{X})] + \min_{\vec{c}\in\mathbb{R}^k}\{\vec{c}\cdot I(\vec{\theta})\cdot(\vec{c})^t - 2\vec{c}\cdot\gamma(\vec{\theta})^t\}.$$

Ora, la matrice di informazione di Fisher è per ipotesi invertibile, sicché, grazie alla precedente proposizione, essa è quindi definita positiva. Possiamo allora applicare il lemma, secondo cui il minimizzatore è:

$$\hat{c} := \arg\Big\{\min_{\vec{c}\in\mathbb{R}^k}\{\vec{c}\cdot I(\vec{\theta})\cdot(\vec{c})^t - 2\vec{c}\cdot\gamma(\vec{\theta})^t\}\Big\} = \gamma(\vec{\theta})\cdot I^{-1}(\vec{\theta}).$$

Sostituendo allora $\hat{c}$ nella varianza e ricordando che $I(\vec{\theta})$ è simmetrica (e quindi anche la sua inversa) si ottiene infine la tesi:

$$\begin{aligned}
0 \le \min_{\vec{c}\in\mathbb{R}^k}\{\mathrm{Var}[T(\vec{X}) - W(\vec{X},\vec{\theta})]\} &= \mathrm{Var}[T(\vec{X})] + \hat{c}\cdot I(\vec{\theta})\cdot(\hat{c})^t - 2\hat{c}\cdot\gamma(\vec{\theta})^t \\
&= \mathrm{Var}[T(\vec{X})] - \gamma(\vec{\theta})\cdot I^{-1}(\vec{\theta})\cdot\gamma(\vec{\theta})^t.
\end{aligned}$$

□

Corollario. *Nelle ipotesi del teorema di Rao-Cramér, se* $\forall 1 \leq i, j \leq k$ *abbiamo che* $T_j = T_j(\vec{X})$ *è uno stimatore non distorto del parametro* θ_j*, allora* $\mathrm{Var}[T_j(\vec{X})] \geq I_{jj}^{-1}(\vec{\theta})$.

Esempio: Sia $\vec{X} \coloneqq (X_1, \ldots, X_n) \sim N(\mu, \sigma^2)$ un campione casuale. Allora

$$L(\mu, \sigma^2, \vec{x}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x_1-\mu)^2} = (2\pi\sigma^2)^{-\frac{n}{2}} e^{\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2}$$

La funzione di log-verosimiglianza sarà allora

$$l(\mu, \sigma^2, \vec{x}) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2.$$

Facilmente possiamo calcolare la matrice di informazione di Fisher, che risulterà:

$$I(\vec{\theta}) = \begin{pmatrix} \frac{n}{\sigma^2} & 0 \\ 0 & \frac{n}{2\sigma^4} \end{pmatrix}, \text{ e quindi } \left(I(\vec{\theta})\right)^{-1} = \begin{pmatrix} \frac{\sigma^2}{n} & 0 \\ 0 & \frac{2\sigma^4}{n} \end{pmatrix}$$

Osserviamo anche che $\mathrm{Var}[\overline{X}_n] = \frac{\sigma^2}{n}$ e $\mathrm{Var}[S_n^2] = \frac{2\sigma^4}{n-1} > \frac{2\sigma^4}{n}$, dove sono stati calcolati i valori dell'inversa della matrice di Fisher. Si deduce allora dal teorema di Rao-Cramér che la media campionaria è stimatore efficiente per μ, mentre non lo è la varianza campionaria per σ^2, anche se lo è asintoticamente. Coerentemente, gli stimatori di massima verosimiglianza si possono ottenere risolvendo il sistema composto dalle rispettive score-functions dei parametri,

$$\begin{cases} \frac{\partial}{\partial\mu} l(\mu, \sigma^2, \vec{X}) = 0 \\ \frac{\partial}{\partial\sigma^2} l(\mu, \sigma^2, \vec{X}) = 0 \end{cases}$$

che ci restituisce le due soluzioni $\hat{\mu}_n = \overline{X}_n$ e $\hat{\sigma^2}_n = \frac{n}{n-1} S_n^2$. Notiamo infine che i due valori non diagonali della matrice di Fisher sono nulli perché ci troviamo in distribuzione normale, ove la non correlazione implica anche l'indipendenza.

Osservazione: Sia $Z \sim N(0, 1)$ una variabile casuale. Allora vale la seguente relazione:

$$\mu_{2s} \coloneqq \mathbb{E}[Z^{2s}] = \frac{(2s)!}{2^s s!}$$

Essa risulta comoda per il calcolo dei momenti delle variabili normali, tenendo conto che $X \coloneqq \mu + \sigma Z \Longrightarrow X \sim N(\mu, \sigma^2)$.

Esempio: Sia $\vec{X} \coloneqq (X_1, \ldots, X_n) \sim f_X(x, \eta) \coloneqq \eta e^{-\eta(x-3)} \mathbb{1}_{[3,+\infty]}$, con $\eta > 0$. Vogliamo calcolare il limite inferiore di Rao-Cramér per uno stimatore non distorto di $g(\eta) \coloneqq \frac{1}{\eta}$, individuare possibilmente un siffatto stimatore e, dopo averlo trovato, calcolare se esso sia o meno efficiente. In base al teorema di Rao-Cramér si ha che

$$I(g(\eta)) = \left(g'(\eta)\right)^2 \frac{1}{I(\eta)} = \frac{1}{\eta^4}\frac{1}{I(\eta)}.$$

Per calcolare $I(\eta)$, troviamo innanzi tutto la funzione di log-verosimiglianza. Si ha per prima cosa che:

$$L(\eta, \vec{x}) = \eta^n e^{-\eta \sum_{i=1}^n (x_i - 3)} \Rightarrow l(\eta, \vec{x}) = n \log(\eta) - \eta \sum_{i=1}^n (x_i - 3)$$

Possiamo calcolare adesso

$$I(\eta) = -\mathbb{E}_\eta \Big[\frac{\partial^2}{\partial \eta^2} l(\eta, \vec{x}) \Big] = -\mathbb{E}_\eta [-\frac{n}{\eta^2}] = \frac{n}{\eta^2}.$$

Ne segue subito che il limite inferiore cercato sarà allora

$$I(g(\eta)) = \frac{1}{\eta^4} \cdot \frac{\eta^2}{n} = \frac{1}{n\eta^2}.$$

Per trovare un possibile stimatore, ricordiamo la relazione già dimostrata durante la dimostrazione del teorema di Rao-Cramér $\mathbb{E}[S(\eta)] = 0$. Ne segue che $0 = \mathbb{E}_\eta \Big[\frac{\partial}{\partial \eta} l(\eta, \vec{X}) \Big] = \mathbb{E}_\eta [\frac{n}{\eta} - \sum_{i=1}^n (X_i - 3)] = n \mathbb{E}_\eta [\frac{1}{\eta} - (\overline{X}_n - 3)] = n \left(\frac{1}{\eta} - \mathbb{E}_\eta [\overline{X}_n - 3] \right)$ da cui

$$T_n(\vec{X}) \coloneqq (\overline{X}_n - 3)$$

è lo stimatore cercato. Ora, è semplice calcolare che

$$\mathrm{Var}[T_n] = \mathrm{Var}[\overline{X}_n] = \frac{1}{n^2} \sum_{i=1}^n \mathrm{Var}[X_i] = \frac{1}{n} \mathrm{Var}[X] = \frac{1}{n\eta^2}$$

Ne segue che lo stimatore cercato è effettivamente anche efficiente.

1.7 Sufficienza

Introduzione alla sufficienza (scritta da Michele Nardin)

Rinfreschiamo il concetto di distribuzione condizionata:

Definizione 8. (Caso discreto) Siano X, Y variabili casuali discrete e supponiamo P(X=x) $\neq 0$. La distribuzione condizionata di Y dato X=x è

$$P_{Y|X} := P(Y = y | X = x) = \frac{P(Y = y \cap X = x)}{P(X = x)}$$

(Caso continuo) Siano X, Y variabili casuali definite su $(\Omega, \mathcal{E}, P)$, con funzioni di densità f_X, f_Y rispettivamente, e densità congiunta $f_{X,Y}$. Ovviamente è assurdo richiedere che P(X=x)$\neq 0$. Allora si ragiona tramite densità condizionate: supponendo che $f_X(x) \neq 0$

$$f_{Y|X}(y|X = x) := \frac{f_{X,Y}(x, y)}{f_X(x)}$$

si dice densità condizionata di Y dato X=x

NB: $f_{Y|X}(y|X=x) = f_Y(y)$ per ogni y se e solo se X e Y sono indipendenti.

La sufficienza di una statistica definisce formalmente la capacità di tale funzione di rappresentare in maniera *sintetica* l'informazione contenuta nel campione, senza perdita di *informazione rilevante*.

Intuitivamente si può pensare alla proprietà di conservazione dell'informazione rilevante in questo modo: qualsiasi altra statistica, calcolata a partire dallo stesso campione, non porta più informazioni rispetto a ϑ di quante ne abbia già portato la statistica sufficiente. Per questo, banalmente, l'intero campione è sicuramente una statistica sufficiente (molto poco sintetica, ma a volte non c'è di meglio).

Sia $(X_1, X_2, ..., X_n)$ un campione casuale da una distribuzione avente funzione di ripartizione $F_X(x, \vartheta)$. Sia inoltre $T_n(X_1, ..., X_n)$ uno stimatore di ϑ avente funzione di ripartizione $F_{T_n}(t, \vartheta)$.

Definizione 9. La statistica T_n viene detta sufficiente per ϑ se la distribuzione di $X_1, ..., X_n$ condizionata a $T_n = t_n$ non dipende da ϑ.

Abbiamo quindi formalizzato le richieste fatte sopra: il fatto che la distribuzione condizionata (del vettore allo stimatore) non dipenda da ϑ implica di fatto l'impossibilità di perdere informazioni rilevanti su ϑ stesso, e cioè, lo stimatore T_n contiene tutte le informazioni necessarie riguardanti ϑ per fare inferenza sul parametro incognito.
In altre parole, *le informazioni contenute in* T_n riguardo a ϑ *sono le stesse contenute nell'intero campione.*
Vedremo che la sufficienza può essere verificata usando il teorema di fattorizzazione di Neyman, che di fatto fornisce una caratterizzazione 'più comoda' da verificare (rispetto all'uso brutale della definizione) per assicurarsi che una data statistica abbia la proprietà desiderata.

Sufficienza (Lezione del 06/05, Scritta da Marco Peruzzetto)

Definizione. Sia $\vec{X} \coloneqq (X_1, \ldots, X_n)$ un campione casuale avente densità $f_X(x, \theta)$ e sia T_n una statistica. Allora T_n è detta *Statistica Sufficiente* per θ se la densita del vettore $\vec{X}$ condizionata a $T_n(\vec{X}) = t_n$ non dipende da θ.

Esempi:

1. Sia $\vec{X} = (X_1, \ldots, X_n) \sim b(1, p)$. Allora $T_n(\vec{X}) \coloneqq \sum_{i=1}^n X_i$ è una statistica sufficiente per p. Infatti, in base alla definizione di statistica sufficiente, si ottiene che

$$f_{\vec{X}|T_n(\vec{X})}(\vec{x}, t_n) \coloneqq \frac{f_{(\vec{X}, T_n(\vec{X}))}(\vec{x}, t_n)}{f_{T_n(\vec{X})}(t_n)} = \frac{p^{t_n}(1-p)^{n-t_n}}{\binom{n}{t_n} p^{t_n}(1-p)n - t_n} = \frac{1}{\binom{n}{t_n}}$$

   e dunque non dipende dal parametro θ.

2. Sia $\vec{X} \coloneqq (X_1, \ldots, X_n) \sim G(\alpha = 2, \beta)$. Allora $T_n(\vec{X}) \coloneqq \sum_{i=1}^n X_i$ è sufficiente per β. Infatti, si ha innanzi tutto che $T_n(\vec{X}) \sim G(2n, \beta)$ grazie all'indipendenza di $\vec{X}$ e alla

proprietà di riproducibilità di G. Allora

$$\begin{aligned} f_{\vec{X}|T_n(\vec{X})}(\vec{x}, t_n) &:= \frac{f_{(\vec{X},T_n(\vec{X}))}(\vec{x}, t_n)}{f_{T_n(\vec{X})}(t_n)} = \frac{f_{\vec{X}}(\vec{x})}{f_{T_n(\vec{X})}(t_n)} \\ &= \frac{\prod_{i=1}^n \frac{1}{\Gamma(2)\beta^2} x_i^{2-1} e^{-\frac{1}{\beta}x_i}}{\frac{1}{\Gamma(2n)\beta^{2n}} t_n^{2n-1} e^{-\frac{1}{\beta}t_n}} \end{aligned}$$

e si vede quindi che si semplificano tutti i termini in β, per cui è sufficiente.

3. Sia $\vec{X} := (X_1, \ldots, X_n)$ un campione casuale avente funzione di densità $f_X(x, \theta) := e^{-(x-\theta)} \mathbb{1}_{(\theta,+\infty)}(x)$, $\theta > 0$. Allora la statistica d'ordine $X_{(1)}$ è sufficiente per θ. Infatti:

$$f_{X_{(1)}}(x, \theta) = ne^{-n(x-\theta)}$$

ed otteniamo subito che

$$f_{\vec{X}|X_{(1)}}(\vec{x}, x_{(1)}, \theta) = \frac{f_{\vec{X}}(\vec{x}, \theta)}{f_{X_{(1)}}(x_{(1)})} = \frac{\prod_{i=1}^n e^{-(x_i-\theta)} \mathbb{1}_{(\theta,+\infty)}(x_i)}{ne^{-n(x_{(1)}-\theta)} \mathbb{1}_{(\theta,+\infty)}(x_{(1)})} = \frac{e^{-\sum_{i=1}^n x_i}}{ne^{-nx_{(1)}}}$$

che non dipende dal parametro θ.

Teorema 17 (di Fattorizzazione (Neyman))**.** *Sia $\vec{X} := (X_1, \ldots, X_n)$ un campione casuale avente densità $f_X(x, \theta)$. Allora $T_n = T_n(\vec{X})$ è statistica sufficiente per θ se e solo se esistono due funzioni non negative g, h, con*

(·) *$h = h(\vec{x})$ e NON dipende da θ*

(·) *$g = g(\theta, t_n(\vec{x}))$ dipende da θ e da $(X_1, \ldots, X_n)$ solo attraverso t_n*

tali che $\forall \vec{x}_0 \in \mathfrak{X}$ e $\forall \theta_0 \in \Theta$, la funzione di verosimiglianza possa essere scritta come

$$L(\theta_0, \vec{x}_0) = h(\vec{x}_0) \cdot g(\theta_0, t_n(\vec{x}_0))$$

Dimostrazione. [⇒]. Sia ha:

$$L(\theta, \vec{x}) = f_{\vec{X}|T_n=t_n}(\vec{x}, t_n, \theta) \cdot f_{T_n}(t_n, \theta) = f_{\vec{X}|T_n=t_n}(\vec{x}, t_n) \cdot f_{T_n}(t_n, \theta),$$

dove la prima uguaglianza viene dalla definizione di densità condizionata, mentre la seconda viene dalla sufficienza di T_n, per cui la densità condizionata non dipende dal parametro θ. Possiamo scegliere allora $h = f_{\vec{X}|T_n=t_n}$ e $g = f_{T_n}$.
[⇐]. Dimostreremo qui il caso in cui il campione casuale sia composto da variabili discrete. Sia quindi $L(\theta, \vec{x}) = h(\vec{x}) \cdot g(\theta, t_n(\vec{x}))$ e definiamo il nuovo insieme $A_{t_n} := \{\vec{x} \in \mathfrak{X} : T_n(\vec{x}) = t_n\}$. Allora

$$f_{T_n}(t_n) = \sum_{\vec{x} \in A_{t_n}} L(\theta, \vec{x}) = \sum_{\vec{x} \in A_{t_n}} h(\vec{x}) \cdot g(\theta, t_n(\vec{x})) = g(\theta, t_n(\vec{x})) \cdot \sum_{\vec{x} \in A_{t_n}} h(\vec{x}),$$

perciò, per definizione di densità condizionata si ottiene che:

$$f_{\vec{X}|T_n=t_n}(\vec{x}, t_n) = \frac{L(\theta, \vec{x})}{f_{T_n}(t_n)} = \frac{h(\vec{x}) \cdot g(\theta, t_n(\vec{x}))}{g(\theta, t_n(\vec{x})) \cdot \sum_{\vec{x} \in A_{t_n}} h(\vec{x})} = \frac{h(\vec{x})}{\sum_{\vec{x} \in A_{t_n}} h(\vec{x})}.$$

Essa non dipende quindi dal parametro θ, sicché è sufficiente. La dimostrazione per il caso continuo è analoga.

□

Attenzione: da questo punto in poi, le dispense sono state scritte durante la sessione d'esame. Per questo motivo, gli autori non hanno potuto dedicarvi il tempo e l'attenzione che un lavoro di questo tipo richiede. Crediamo che possano essere comunque più precise e complete degli appunti presi in classe, ma non garantiamo nulla. Appena possibile le dispense verranno ricontrollate e rese definitive.

Lezione del 10/05, ultima modifica 10/06, Andrea Gadotti

Il teorema di fattorizzazione costituisce un criterio per l'individuazione, se esiste, di una statistica sufficiente.

Esempio Sia $(X_1, ..., X_n)$ da $N(\mu, \sigma^2)$.

(a) se μ è noto, allora cerco una statistica sufficiente per σ^2.
$T_n(X_1, ..., X_n) = \sum_{i=1}^{n}(X_i - \mu)^2$

(b) Se σ^2 è noto, allora cerco una statistica sufficiente per μ.

(c) Se μ e σ^2 non sono noti, allora cerco una statistica congiuntamente sufficiente per μ e σ^2.
$S_n(X_1, ..., X_n) = \left(\sum_{i=1}^{n} X_i, \sum_{i=1}^{n}(X_i - \overline{X}_n^2)\right)$.

Risulta quindi evidente che il concetto di statistica sufficiente è *problem dependent.*

Esempio 1. Sia $(X_1, ..., X_n)$ da $U(\theta, \theta+1)$ con $\theta \in \mathbb{R}$. Vogliamo trovare una statistica sufficiente per θ.
Calcoliamo innanzitutto la funzione di verosimiglianza:

$$L(\theta; \underline{x}) = \prod_{i=1}^{n} \mathbb{1}_{[\theta;\theta+1]}(x_i) \tag{1.9}$$

$$= \prod_{i=1}^{n} \mathbb{1}_{\mathbb{R}}(x_i)\mathbb{1}_{[\theta;\theta+1]}(x_{(n)})\mathbb{1}_{[\theta;\theta+1]}(x_{(1)}) \tag{1.10}$$

$$= \prod_{i=1}^{n} \mathbb{1}_{\mathbb{R}}(x_i)\mathbb{1}_{[X_{(n)}-1;X_{(n)}]}(\theta) \tag{1.11}$$

Notiamo che l'ultima espressione è il prodotto di due funzioni dove la prima dipende solo dal campione, mentre la seconda dipende sia dal parametro θ che dal campione. Quindi grazie al teorema di fattorizzazione di Neyman abbiamo che $(X_{(1)}, X_{(n)})$ è statistica sufficiente per U.
Nota: poiché $U_{(\theta,\theta+1)}$ non appartiene alla famiglia regolare non è necessariamente vero che *dimensione statistica sufficiente = dimensione vettore parametri.* Infatti in questo caso abbiamo $2 \neq 1$.

Osservazione 9. Spesso vediamo il campione casuale come diverso da una statistica, che è funzione del campione. Ma esso porta in sè tutta l'informazione disponibile relativamente al vettore parametrico. Anche il campione stesso è una statistica, in particolare è una statistica sufficiente, l'unico problema è che non è per nulla "sintetico".

Problema: esiste una statistica sufficiente che sia migliore (ovvero più sintetica) delle altre, a parità di informazione contenuta circa θ? In effetti una tale statistica esiste e viene detta *statistica sufficiente minimale*.

1.7.1 Statistiche sufficienti minimali

Esempio introduttivo Pensiamo di replicare $n = 4$ volte una prova bernoulliana con probabilità di successo p. Consideriamo lo spazio campionario $\varkappa$, composto da $2^4 = 16$ elementi. Vogliamo fare inferenza su p. Definiamo di seguito tre statistiche differenti:

1) $Y_1 :=$ risultato della prima prova
2) $Y_2 :=$ numero di successi nelle quattro prove
3) $Y_3 := (Y_1, Y_2)$

A questo punto notiamo che:

1) Y_1 non è una statistica sufficiente per p (il fatto che ci sia stato un successo o meno nella prima prova non ci dà alcuna informazione rispetto a p)
2) Y_2 è statistica sufficiente per p (cfr. Teorema di fattorizzazione) e consiste in un unico valore, ovvero la somma degli elementi del campione
3) Y_3 consiste in due valori, ovvero è un vettore di dimensione 2. È statistica sufficiente per p, ma è eccessivamente raffinata per il problema che vogliamo affrontare, poiché l'apporto di Y_1 è inutile. Infatti Y_3 non è statistica sufficiente minimale per p.

Definizione 10. Una statistica $T_n(X_1, ..., X_n)$ è detta *statistica sufficiente minimale* per θ se:

1) è statistica sufficiente per θ
2) assume valori distinti solamente in punti dello spazio campionario $\varkappa$ a cui corrispondono verosimiglianze non equivalenti, ovvero se $\forall \underline{x_1}, \underline{x_2} \in \varkappa$ vale

$$T_n(\underline{x_1}) \neq T_n(\underline{x_2}) \iff L(\theta, \underline{x_1}) \neq L(\theta, \underline{x_2}) \ \ \forall \theta \in \Theta$$

Teorema 18 (di Lehmann-Scheffé)**.** *Sia $(X_1, ..., X_n)$ un campione casuale da una distribuzione di densità $f(x; \theta)$ e sia $S_n = S_n(X_1, ..., X_n)$ tale che, prese le determinazioni campionarie $\underline{x}$ e $\underline{y}$, il rapporto tra le funzioni di verosimiglianza valutate in $\underline{x}$ e $\underline{y}$, è funzione che non dipende da θ se e solo se $S_n(\underline{x}) = S_n(\underline{y})$. Allora $S_n(X_1, ..., X_n)$ è statistica sufficiente minimale per θ. Ovvero:*

$$\frac{L(\theta; \underline{x})}{L(\theta; \underline{y})} = m(\underline{x}, \underline{y}) \iff S_n(\underline{x}) = S_n(\underline{y})$$

$\implies S_n(X_1, ..., X_n)$ *è statistica sufficiente minimale per θ.*

Esempio 2. Sia $(X_1, ..., X_n)$ da $X \sim \Gamma(\alpha, \beta), \alpha, \beta > 0$. Vogliamo trovare una statistica congiuntamente sufficiente e minimale per α e β:

1) Cerchiamo una statistica sufficiente:

$$L(\alpha, \beta; \underline{x}) = \left((\Gamma(\alpha)\beta^{\alpha})^{-n} \left(e^{-\frac{1}{\beta}\sum_{i=1}^{n} x_i} \right) \left(\prod_{i=1}^{n} x_i \right) \right) \cdot \left(\prod_{i=1}^{n} \mathbb{1}_{\mathbb{R}_+}(x_i) \right)$$

Osserviamo che abbiamo scritto $L(\alpha, \beta; \underline{x})$ come $g\left(\alpha, \beta; \sum_{i=1}^{n} x_i, \prod_{i=1}^{n} x_i\right) \cdot h(\underline{x})$. Quindi per il teorema di fattorizzazione abbiamo che $S_n := (\sum_{i=1}^{n} x_i, \prod_{i=1}^{n} x_i)$ è statistica congiuntamente sufficiente per α e β.

2) Verifichiamo ora che tale statistica è anche minimale. Supponiamo che $S_n(\underline{x}) = S_n(\underline{y})$, ovvero $(\sum_{i=1}^{n} x_i, \prod_{i=1}^{n} x_i) = (\sum_{i=1}^{n} y_i, \prod_{i=1}^{n} y_i)$. Allora:

$$\frac{L(\alpha, \beta; \underline{x})}{L(\alpha, \beta; \underline{y})} = e^{\frac{1}{\beta}(\sum_i x_i - \sum_i y_i)} \left(\prod_i \frac{x_i}{y_i} \right) = e^0 \left(\prod_i (1) \right)^{\alpha} = 1$$

Per il teorema di Lehmann-Scheffé concludiamo che S_n è statistica sufficiente minimale.
Nota: in realtà noi abbiamo verificato che vale solo una delle due implicazioni richieste dalle ipotesi del teorema (quella da destra a sinistra). Questo è ciò che è stato fatto a lezione, e per il momento scriviamo solo questo.

1.7.2 Principio di verosimiglianza

La verosimiglianza combina due tipi di informazione:

- informazione pre-sperimentale espressa attraverso il modello statistico scelto per descrivere il fenomeno di indagine

$$\mathfrak{F}_\theta = \left\{ f(x; \theta) : \ \theta \in \Theta \subseteq \mathbb{R}^k \right\}$$

- informazione sperimentale contenuta in quella che è la determinazione campionaria $\underline{x} = (x_1, \ldots, x_n)$

$$L(\theta; \underline{x}) = \prod_{i=1}^{n} f_{X_i}(x_i; \theta)$$

Nota: In generale $L(\theta; \underline{x}) = f_{X_1,\ldots,X_n}(x_1, \ldots, x_n; \theta)$, densità congiunta, ovvero è definita anche nel caso in cui le variabili casuali non siano indipendenti ed equidistribuite. Quando invece lo sono, possiamo esprimere $L(\theta; \underline{x})$ nella solita forma, ovvero come produttoria.

Principio di verosimiglianza Con riferimento ad un dato modello stocastico $\mathfrak{F}_\theta = \left\{ f(x; \theta) : \ \theta \in \Theta \subseteq \mathbb{R}^k \right\}$, due punti $\underline{x_1}$ e $\underline{x_2} \in \varkappa$ tali che $L(\theta; \underline{x_2}) = L(\theta; \underline{x_2})$ devono condurre alle medesime conclusioni inferenziali circa θ.

Teorema 19. *L'informazione di Fisher fornita da uno stimatore T_n basato su una statistica sufficiente W_n coincide con l'informazione fornita dall'intero campione*

Dimostrazione. Dal teorema di fattorizzazione l'informazione di Fisher sarà funzione del campione solo attraverso

$$\frac{d}{d\theta}\ln\left[g(W_n);\theta\right] \qquad W_n = T_n(\underline{x})$$

dal momento che la derivata rispetto a θ di h è nulla ovvero $\frac{d}{d\theta}\ln[h] = 0$, dunque

$$I(\theta) = \mathbb{E}_\theta\left[\frac{d}{d\theta}\ln(g(W_n);\theta)\right]^2$$

dove il membro di sinistra è l'informazione associata all'intero campione e il membro di destra è l'informazione di Fisher restituita dalla statistica sufficiente W_n. Questo implica che $I(W_n)$ è ugualmente rappresentativo per l'informazione di Fisher di quanto non lo sia l'intero campione. □

Esempio 3. Sia $(X_1,\ldots,X_n)$ da $N(\mu,\sigma^2=1)$, vogliamo individuare una statistica sufficiente minimale per μ e la relativa informazione di Fisher.

1) Notiamo che
$$W_n := \sum_{i=1}^n X_i \sim N(n\mu, n)$$
è statistica sufficiente (si verifica subito usando il teorema di fattorizzazione). Vogliamo mostrare che l'informazione di Fisher che si ottiene usando solo W_n è la stessa che si trova usando l'intero campione.

2) Calcoliamo $I(\mu)$ supponendo di avere in mano, anziché l'intero campione $(x_1,...,x_n)$, il solo valore $w_n = \sum_{i=1}^n x_i$ (di cui conosciamo la distribuzione, come visto al punto 1). Iniziamo calcolando la funzione di verosimiglianza di w_n (che coincide con la sua densità, essendo un unico valore e non un vettore):
$$L(\mu; w_n) = \frac{1}{\sqrt{2\pi n}}e^{-\frac{1}{2n}(w_n-n\mu)^2}$$
A questo punto, calcolando la funzione di log-verosimiglianza, con un po' di conti si trova facilmente che $I_{W_n}(\mu) = n$.

3) Calcoliamo $I(\mu)$ supponendo di avere in mano l'intero campione. Iniziamo trovando la funzione di verosimiglianza di $(x_1,...,x_n)$:
$$L(\mu;\underline{x}) = \prod_{i=1}^n\left(\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}(x_i-\mu)^2}\right)$$
Come prima, con un po' di conti si ottiene $I_{\underline{x}}(\mu) = n$, ovvero le due informazioni di Fisher coincidono.

Lezione del 13/05, ultima modifica 16/06, Andrea Gadotti

1.7.3 Famiglie esponenziali e sufficienza

Notiamo innanzitutto che una qualsiasi funzione di densità per una distribuzione appartenente alla famiglia esponenziale può essere scritta come

$$f_X(x;\theta) = \exp\left\{ C(x) + D(\theta) \sum_{m=1}^{k} A_m(\theta) B_m(x) \right\}$$

Teorema 20. *Sia $(X_1, ..., X_n)$ un campione casuale appartenente alla famiglia esponenziale a k parametri. Allora il vettore $T_n := (\sum_{i=1}^n B_1(x_i), ..., \sum_{i=1}^n B_k(x_i))$ è statistica sufficiente per θ.*

Dimostrazione. Da inserire (comunque è facile, basta usare il teorema di fattorizzazione). □

Esempio 4. Sia $(X_1, ...X_n)$ da $N(\mu, \sigma^2)$ con μ e σ^2 non noti (dunque $\vec{X}$ proviene da una famiglia esponenziale a 2 parametri). Per il teorema appena visto si ha che $I_n := (\sum_{i=1}^n B_1(x_i), \sum_{i=1}^n B_2(x_i)) = (\sum_{i=1}^n x_i, \sum_{i=1}^n x_i^2$ è statistica sufficiente per (μ, σ^2). (Nota: la seconda uguaglianza si verifica immediatamente).
Affermiamo (senza dimostrazione) che lo stimatore di massima verosimiglianza per (μ, σ^2) è $\hat{\theta} = \left(\frac{1}{n}\sum_{i=1}^n X_i, \frac{1}{n}\sum_{i=1}^n (X_i - \overline{X}_n)^2\right)$ e notiamo che esso è funzione della statistica sufficiente I_n.

Lezione del 17/05, ultima modifica 21/06, Andrea Gadotti

1.7.4 Teorema di Rao-Blackwell

In questa sezione ci occuperemo del seguente teorema molto importante:

Teorema 21 (di Rao-Blackwell)**.** *Sia $(X_1, ..., X_n)$ un campione casuale da una distribuzione avente funzione di densità $f(x;\theta)$. Siano inoltre V_n uno stimatore non distorto di θ e T_n una statistica sufficiente per θ. Allora*

$$V_{n,T_n} = \mathbb{E}_\theta(V_n | T_n)$$

ha le seguenti proprietà:

(1) $\mathbb{E}_\theta(V_{n,T_n}) = \theta, \forall\theta \in \Theta$*, ovvero V_{n,T_n} è uno stimatore non distorto di θ*

(2) $\mathrm{Var}_\theta(V_{n,T_n}) \leq \mathrm{Var}_\theta(V_n)$

(3) $V_{n,T_n} = \phi(T_n)$*, ovvero V_{n,T_n} è funzione della statistica sufficiente T_n.*

Possiamo riassumere l'enunciato dicendo il condizionamento di uno stimatore rispetto a una statistica sufficiente preserva la non distorsione e migliora lo stimatore sotto il profilo della varianza.

Dimostrazione.

(1) Utilizziamo senza dimostrazione il seguente risultato:

$$\mathbb{E}[\mathbb{E}(X|Y)] = \mathbb{E}(X)$$

A questo punto il primo punto è subito dimostrato. Infatti

$$\mathbb{E}_\theta(V_{n,T_n}) = \mathbb{E}_\theta[\mathbb{E}_\theta(V_n|T_n)] = \theta$$

(2) Utilizziamo senza dimostrazione il seguente risultato:

$$\mathrm{Var}(X) = \mathrm{Var}[\mathbb{E}(X|Y)] + \mathbb{E}[\mathrm{Var}(X|Y)]$$

A questo punto anche il secondo punto è subito dimostrato. Infatti

$$\mathrm{Var}(V_n) = \mathrm{Var}[\mathbb{E}(V_n|T_n)] + \mathbb{E}[\mathrm{Var}(V_n|T_n)] = \mathrm{Var}(V_{n,T_n}) + \mathbb{E}[\mathrm{Var}(V_n|T_n)]$$

e si conclude osservando che $\mathbb{E}[\mathrm{Var}(V_n|T_n)] \geq 0$ in quanto valore di aspettazione di una quantità sicuramente non negativa.

(3)

$$V_{n,T_n} = \mathbb{E}_\theta(V_n|T_n) = \int_D v_n \cdot f_{v_n|t_n}(v_n, t_n)\, dv_n = \int_{D^n} v_n(x_1, ..., x_n) \cdot f_{\underline{x}|t_n}(\underline{x}, t_n)\, d\underline{x} = \phi(T_n)$$

dove l'ultima uguaglianza è giustificata dal fatto che $f_{\underline{x}|t_n}(\underline{x}, t_n)$ non dipende da θ per definizione di statistica sufficiente.

□

Esempio 5. Sia $(X_1, ..., X_n)$ da $b(1, p), p \in (0, 1)$. Sappiamo che $T_n = \sum_{i=1}^n X_i$ è statistica sufficiente minimale per p. Vogliamo applicare il teorema di Rao-Blackwell con lo stimatore X_1. Notiamo che lo stimatore scelto è estremamente "grezzo", dato che il primo risultato di n prove non ci dice in realtà nulla su p, ma nonostante questo lo stimatore che otterremo grazie al teorema sarà molto buono.

a) Notiamo innanzitutto che $V_n := X_1$ è stimatore non distorto per p. Infatti $\mathbb{E}(V_n) = \mathbb{E}(X_1) = p$

b) Vogliamo costruire ora V_{n,T_n}:

$$\begin{aligned}
V_{n,T_n} := \mathbb{E} &= 0 \cdot \mathbb{P}(X_1 = 0|T_n) + 1 \cdot \mathbb{P}(X_1 = 1|T_n) = \mathbb{P}(X_1 = 1|T_n = t_n) \\
&= \frac{\mathbb{P}(X_1 = 1|T_n = t_n)}{\mathbb{P}(T_n = t_n)} \\
&= \frac{\mathbb{P}(X_1 = 1|T_{n-1} = t_n - 1)}{\mathbb{P}(T_n = t_n)} \\
&= \frac{\mathbb{P}(X_1 = 1)\mathbb{P}(\sum_{i=2}^{n} x_i = t_n - 1)}{\mathbb{P}(T_n = t_n)} \\
&= \frac{p\binom{n-1}{t_n-1}p^{t_n-1}(1-p)^{(n-1)-(t_n-1)}}{\binom{n}{t_n}p^{t_n}(1-p)^{n-t_n}} \\
&= t_n/n \\
&= \frac{1}{n}\sum_{i=1}^{n} x_i = \overline{x}_n
\end{aligned}$$

dove abbiamo usato anche il fatto che $T_{n-1} = \sum_{i=2}^{n} x_i$ e $T_n \sim b(n, p)$.

c) Per il teorema di Rao-Blackwell siamo sicuri di aver trovato uno stimatore *uguale o migliore* (in termini di varianza) di quello iniziale (X_1). Notiamo però che in realtà abbiamo ottenuto il migliore in assoluto, ovvero un UMVU di p. Infatti:

$$I(p) = -\mathbb{E}\left(\frac{d^2}{dp^2} l(p; \underline{x})\right) = \frac{n}{p(1-p)} = \frac{1}{\mathrm{Var}(\overline{X}_n)}$$

e poiché $b(1, p)$ appartiene alla famiglia regolare, possiamo applicare il teorema di Rao-Cramer, e quindi concludere che lo stimatore trovato, ovvero la media campionaria, è quello con varianza uniformemente minima, ovvero è un UMVU.

Osservazione 10.

1) Nell'esempio precedente, per verificare che lo stimatore ha varianza minima abbiamo trovato $I(\theta)$ e $1/\mathrm{Var}(T_n)$ e abbiamo mostrato che sono uguali. Grazie al lemma 1 di pagina 41, potevamo invece trovare $I(\theta)$ e la score function $S(\theta) := \frac{d}{d\theta} l(\theta, \underline{x})$ e mostrare che $S(\theta) = (Vn, T_n - \theta)I(\theta)$ (sempre solo nel caso in cui abbiamo a che fare con una famiglia regolare)

2) Quanto detto nel punto 1) è utile nel caso in cui lo stimatore trovato sia effettivamente efficiente. Purtroppo però non tutti gli stimatori UMVU sono efficienti (ovvero: potrebbe non esistere uno stimatore non distorto che sia efficiente), anche se la distribuzione appartiene alla famiglia regolare.

3) Nel caso di famiglie non regolari, non possiamo utilizzare il teorema di Rao-Cramer. Dobbiamo quindi trovare trovare un altro modo per verificare se, una volta trovato uno stimatore non distorto, questo ha anche varianza uniformemente minima.

4) Il teorema di Rao-Blackwell (in particolare il punto 3) afferma che la ricerca di uno stimatore ottimale si può restringere alla classe degli stimatori non distorti che siano funzioni di statistiche sufficienti per il parametro su cui si vuole fare inferenza. Finora abbiamo trattato il concetto di sufficienza. Come vedremo, quest'ultima, unita alla *completezza*, ci permetterà di risolvere il problema della ricerca di uno stimatore ottimale nel caso in cui il campione non provenga da una famiglia regolare.

1.8 Completezza

Esiste la possibilità di avere l'unicità dello stimatore UMVU, ovvero l'unicià dello stimatore di minima varianza, all'interno della classe degli stimatori non distorti. A tal fine, introduciamo il concetto di *completezza*.

Definizione 11. Una statistica T_n è detta *completa* per la famiglia di distribuzioni da cui proviene se per qualunque funzione $\phi(T_n)$ vale la seguente implicazione:

$$\mathbb{E}_\theta[\phi(T_n)] = 0 \ \ \forall\theta \in \Theta \Longrightarrow \mathbb{P}(\phi(T_n) = 0) = 1, \text{ ovvero } \phi(T_n) = 0 \text{ quasi ovunque}$$

1.8.1 Teorema di Lehmann-Scheffé

Teorema 22 (di Lehmann-Scheffé)**.** *Sia T_n una statistica sufficiente e completa per θ e sia $V_n(X_1, ..., X_n)$ uno stimatore non distorto per θ. Allora*

$$\phi(T_n) = Vn, T_n = \mathbb{E}_\theta(V_n|T_n)$$

è l'unica funzione della statistica sufficiente T_n che risulta essere stimatore non distorto di θ.

Dimostrazione. Sia $\phi^*(T_n)$ un'altra funzione di T_n, anch'essa non distorta per θ. Allora $\mathbb{E}_\theta[\phi(T_n) - \phi^*(T_n)] = \mathbb{E}_\theta[\phi(T_n)] - \mathbb{E}_\theta[\phi^*(T_n)] = \theta - \theta = 0$.
Dalla completezza di T_n si ha che

$$\mathbb{E}_\theta[\phi(T_n) - \phi^*(T_n)] = 0 \Longrightarrow \phi(T_n) - \phi^*(T_n) = 0 \text{ q. o.} \Longrightarrow \phi(T_n) = \phi^*(T_n) \text{ q. o.}$$

□

Osservazione 11 (importante). Grazie al teorema di Rao-Blackwell sapevamo che possiamo restringere la ricerca di uno stimatore UMVU alle funzioni di statistiche sufficienti. Grazie al teorema di Lehmann-Scheffé possiamo ora dire che (se la statistica è completa), la funzione migliore da considerare è unica ed è proprio $\mathbb{E}_\theta(V_n|T_n)$. Notiamo che se la statistica sufficiente è più di una, possiamo sempre ricondizionare alla statistica T_n qualsiasi stimatore ottenuto usando le altre statistiche, ottenendo una funzione di T_n. Per questo possiamo dire che $\mathbb{E}_\theta(V_n|T_n)$ è l'unico stimatore UMVU per θ.

Problema: abbiamo visto come la completezza di una statistica sufficiente ci aiuta nella ricerca di uno stimatore UMVU. Ma la completezza è una proprietà ricorrente?

Teorema 23. *Data una famiglia esponenziale a k parametri, il vettore*

$$\left(\sum_{i=1}^{n} B_1(x_i), ..., \sum_{i=1}^{n} B_k(x_i)\right)$$

è statistica congiuntamente sufficiente (e minimale) per il vettore dei parametri $\underline{\theta} = (\theta_1, ..., \theta_k)$*, ed è anche statistica completa.*

Esempio 6 (riassuntivo). Sia $(X_1, ..., X_n)$ un campione casuale da distribuzione avente densità esponenziale

$$f_X(x;\theta) = \theta e^{-\theta x} \mathbb{1}_{\mathbb{R}^+}(x), \quad \text{con } \theta > 0$$

Vogliamo costruire uno stimatore UMVU per θ.

a) Cerchiamo una statistica sufficiente per θ:

$$L(\theta, \underline{x}) = \prod_{i=1}^{n} \theta e^{-\theta x_i} \mathbb{1}_{\mathbb{R}^+}(x_i) = \theta^n e^{-\theta \sum_{i=1}^{n} x_i} \prod_{i=1}^{n} \mathbb{1}_{\mathbb{R}^+}(x_i)$$

Per il teorema di fattorizzazione di Neymann, $T_n := \sum_{i=1}^{n} X_i$ è statistica sufficiente per θ.
Notiamo che, poiché la distribuzione in questione appartiene alla famiglia esponenziale, la statistica trovata è anche completa.

b) Cerchiamo lo stimatore di massima verosimiglianza per θ.

$$l(\theta, \underline{x}) = n \log(\theta) - \theta \sum_{i=1}^{n} x_i + \sum_{i=1}^{n} \log(x_i)$$

$$\frac{d}{d\theta} l(\theta, \underline{x}) = n/\theta - \sum_{i=1}^{n} x_i$$

Ponendo la derivata uguale a 0 si ottiene subito che $\hat{\theta}_n := \frac{n}{\sum_{i=1}^{n} X_i}$ è stimatore di massima verosimiglianza per θ.

c) Vediamo se $\hat{\theta}_n$ è non distorto.
Osserviamo innanzitutto che $X_i \sim \gamma(\alpha = 1, \beta = 1/\theta)$. Dunque

$$W := \sum_{i=1}^{n} x_i \sim \gamma(\alpha = n, \beta = 1/\theta)$$

Quindi:

$$\mathbb{E}(\hat{\theta}_n) = \mathbb{E}\left(\frac{n}{\sum_{i=1}^{n} X_i}\right) = n\mathbb{E}\left(\frac{1}{W}\right) = n \int_0^{+\infty} \frac{1}{w} f_W(w, \alpha, \beta) dw = ... = \frac{n}{n-1}\theta \neq \theta$$

Quindi $\hat{\theta}_n$ è stimatore distorto per θ, ma è curabile facilmente.

d) Consideriamo la statistica di S_n data da $S_n := \frac{n-1}{n}\hat{\theta}_n$. Notiamo che si tratta di uno stimatore non distorto di θ, funzione della statistica sufficiente e completa $\sum_{i=1}^{n} X_i$. Quindi per i teoremi di Rao-Blackwell e Lehmann-Scheffé, S_n è lo stimatore UMVU per θ.

Esempio 7. Sia $(X_1, ..., X_n)$ da $U(0, \theta)$. Notiamo che la distribuzione in questione non appartiene alla famiglia esponenziale (perché il suo supporto dipende da θ).

a) Sappiamo che $X_{(n)} = \max(X_1, ..., X_n)$ è una statistica sufficiente per θ. Si dimostra (esercizio) che è anche completa.

b) Abbiamo visto in precedenza che $\frac{n-1}{n}X_{(n)}$ è stimatore non distorto per θ, ed è chiaramente anche funzione di $X_{(n)}$

c) Quindi per i teoremi di Rao-Blackwell e Lehmann-Scheffé, $\frac{n-1}{n}X_{n)}$ è lo stimatore UMVU per θ.

Lezione del 20/05, ultima modifica 21/06, Andrea Gadotti

1.9 Proprietà degli stimatori di massima verosimiglianza

Gli stimatori di massima verosimiglianza godono delle seguenti proprietà:

(1) **Relazione con statistiche sufficienti:** gli stimatori di MV sono funzioni di statistiche sufficienti. Infatti

$$L(\theta, \underline{x}) = g(t_n(x), \theta)h(x) \Longrightarrow l(\theta, \underline{x}) = \log(g(t_n(x), \theta)) + \log(h(x))$$

e quindi lo stimatore di massima verosimiglianza risulterà funzione della sola t_n.

(2) **Proprietà di invarianza:** se $\hat{\theta}_n$ è stimatore di MV per θ e $g(\theta)$ una funzione di θ, allora $g(\hat{\theta}_n)$ è stimatore di MV per $g(\theta)$.

(3) **Efficienza (per n finito):** nel caso in cui la distribuzione provenga da una famiglia regolare, se esiste uno stimatore non distorto T_n di θ la cui varianza raggiunge il limite inferiore di Rao-Cramer, tale stimatore coincide con lo stimatore di MV di θ. (nota: in generale gli stimatori di MV sono distorti)

(4) **Consistenza:** gli stimatori di MV sono consistenti

- in senso forte (ovvero sono *quadraticamente consistenti*):

$$\mathrm{MSE}_\theta(T_n) = \mathrm{Var}_\theta(T_n) + B_\theta(T_n) \overset{n\to\infty}{\longrightarrow} 0$$

- in senso debole (ovvero sono *semplicemente consistenti*)

$$\mathbb{P}(|\hat{\theta}_n - \theta| > \varepsilon) \overset{n\to\infty}{\longrightarrow} 0$$

(5) **Efficienza (per $n \to \infty$):** sotto condizioni molto generali di regolarità, lo stimatore di MV $\hat{\theta}_n$ è asintoticamente efficiente, ovvero $\hat{\theta}_n$ è tale che:

a) $\lim_{n\to\infty} \mathbb{E}_\theta(\hat{\theta}_n) = \theta \ \ \forall\theta \in \Theta$

b) $\lim_{n\to\infty} \mathrm{Var}_\theta(\hat{\theta}_n) = \frac{1}{I(\theta)} \ \ \forall\theta \in \Theta$

(6) **Distribuzione asintotica:** sempre sotto alcune ipotesi di regolarità, lo stimatore di massima verosimiglianza per n grande ha distribuzione normale, ovvero

$$\hat{\theta}_n \overset{a}{\sim} N\left(\theta, \frac{1}{I(\theta)}\right)$$

Questo risultato risulta molto utile in ambito inferenziale:

Esempio 8.

$$IC_\theta(\alpha) = \left[\hat{\theta}_n - z_{1-\alpha/2}\frac{1}{\sqrt{I(\hat{\theta}_n)}},\ \ \hat{\theta}_n + z_{1-\alpha/2}\frac{1}{\sqrt{I(\hat{\theta}_n)}}\right]$$

dove possiamo anche sfruttare la proprietà $I(\hat{\theta}_n) = nI_1(\hat{\theta}_n)$.

Osservazione 12. Sia $g(\theta)$ una funzione continua di θ e derivabile in θ_0 tale che $g'(\theta_0) \neq 0$. Allora

$$\sqrt{n}(g(\hat{\theta}_n - g(\theta_0)) \xrightarrow{d} N\left(0, \frac{(g'(\theta_o))^2}{I(\theta_0)}\right)$$

dove notiamo che, per la proprietà (1), $g(\hat{\theta}_n)$ è lo stimatore di MV per $g(\theta)$. La dimostrazione è immediata usando il Δ-method.

Lezione 24/05, ultima modifica 01/06, Michele Nardin

Stimatori UMVU per funzioni di ϑ

Sappiamo che lo stimatore di massima verosimiglianza per $\eta(\vartheta)$ è $\eta(\hat{\vartheta})$, ove ovviamente $\hat{\vartheta}$ è stimatore di ML (maximum likelihood) di ϑ. Possiamo affidarci a diverse strategie:

1. Usare *direttamente* il teorema di Rao-Blackwell, e quindi calcolare il valore atteso condizionato di V_n (stimatore non distorto di $\eta(\vartheta)$) a T_n (statistica sufficiente per $\eta(\vartheta)$)
$$V_{n;T_n} = E_\vartheta(V_n|T_n)$$
2. Usare *indirettamente* il teorema di Rao-Blackwell. Sapendo che, se esiste, uno stimatore UMVU per $\eta(\vartheta)$ deve essere funzione di statistica sufficiente, ossia
$$V_{n;T_n} = \phi(T_n)$$

possiamo "aggirare" il problema, ossia studiare il valore atteso di una statistica sufficiente (ricavata in qualche modo): se tale statistica (sufficiente) risulta anche non distorta, per Rao-Blackwell abbiamo lo stimatore UMVU per $\eta(\vartheta)$.

Illustriamo quanto detto al punto 2. con il seguente
Esempio importante: Sia $(X_1, ..., X_n)$ da $b(1, p)$. Sappiamo che $Var(X_i) = p(1-p) =: \eta(p)$. Cerchiamo quindi uno stimatore UMVU per $Var(X_i)$, che è funzione di p. Partiamo da

$$L(p, \vec{X}) = p^{\sum_i X_i}(1-p)^{n-\sum_i X_i} \prod_{i=1}^{n} \mathbb{1}_{0,1}(X_i)$$

usando il teorema di fattorizzazione di Neyman, ci accorgiamo che L può essere scomposta in

$$L(p, \vec{X}) = g(t_n, p)h(\vec{X})$$

ove

$$h(\vec{X}) = \prod_{i=1}^{n} \mathbb{1}_{\{0,1\}}(X_i)$$

e

$$g(t_n, p) = p^{t_n}(1-p)^{n-t_n}$$

dove ovviamente

$$t_n = \sum_{i=1}^{n} X_i$$

Segue che $\hat{p} := \frac{1}{n}t_n$ è statistica sufficiente per p, ed è pure stimatore di ML. (Ma tutto questo già lo sapevamo, è solo per rinfrescare un po' le idee. Adesso?)

Adesso consideriamo $\hat{\eta}(p) := \eta(\hat{p}) = \hat{p}(1-\hat{p})$. Esso, per proprietà di invarianza, è stimatore ML per $\eta(p)$. Segue quindi che è pure funzione di statistiche sufficienti, per la prima proprietà degli stimatori ML. Ma allora, se $E(\hat{\eta}(p)) = Var(X_i)$, per il teorema di Rao-Blackwell ho concluso! Purtroppo $\hat{\eta}(p)$ è solo *asintoticamente* non distorto, infatti: (ricordiamo che $Var(t_n) = E(t_n^2) - E(t_n)^2$)

$$\begin{aligned} E_p(\hat{p}(1-\hat{p})) &= \frac{1}{n^2}E(nt_n - t_n^2) = \frac{1}{n^2}\left\{nE(t_n) - [Var(t_n) + E(t_n)^2]\right\} \\ &= \frac{1}{n^2}\left\{n^2p - np(1-p) - n^2p^2\right\} = \frac{n-1}{n}p(1-p) \\ &\neq p(1-p) \end{aligned}$$

Ma allora, se considero

$$V_{n;t_n} = \frac{n}{n-1}t_n = \frac{n}{n-1}\hat{p}(1-\hat{p})$$

esso è uno stimatore non distorto, e dato che è funzione di statistica sufficiente per $\eta(p) = p(1-p)$ allora per Rao-Blackwell è pure uno stimatore UMVU per $\eta(p)$!

Relazione tra le distribuzioni degli stimatori ML e UMVU per $\eta(p)$

Vogliamo confrontare $\hat{\eta} = \eta(\hat{p}) = \hat{p}(1-\hat{p})$ con $V_{n;t_n} = \frac{n}{n-1}\hat{p}(1-\hat{p})$. Grazie alla consistenza di $\hat{\eta}$ (che discende a sua volta dalla consistenza di $\hat{p}$) vale

$$\hat{\eta} - V_{n;t_n} = \frac{n}{n-1}\hat{\eta} - \hat{\eta} = \frac{1}{n-1}\hat{\eta} \xrightarrow{P} 0$$

Segue quindi che pure $V_{n;t_n}$ è consistente per $\eta(p)$. Inoltre $\hat{\eta}$ e $V_{n;t_n}$ hanno la stessa distribuzione asintotica, in particolare da

$$\sqrt{n}(\hat{\eta} - \eta(p)) - \sqrt{n}(V_{n;t_n} - \eta(p)) = \frac{\sqrt{n}}{n-1}\hat{\eta} \xrightarrow{P} 0$$

e da (uso metodo DELTA, osservando che $\eta(p)$ è funzione continua di p e che $\eta'(p) = 1 - 2p \neq 0$ se $p \neq 1/2$)

$$\sqrt{n}(\hat{\eta} - \eta(p)) \xrightarrow{D} N(0, (1-2p)p(1-p))$$

segue che

$$\sqrt{n}(V_{n;t_n} - \eta(p)) \xrightarrow{D} N(0, (1-2p)p(1-p))$$

Nota: in questo modo abbiamo visto una cosa che vale in generale (davvero?): le conoscenze che abbiamo riguardanti gli stimatori ML possiamo riversarle sugli stimatori UMVU, poichè hanno la stessa distribuzione asintotica!

1.10 Metodi Numerici per la Stima di Massima Verosimiglianza

(da sistemare)

Le equazioni di stima di massima verosimiglianza spesso non sono risolvibili analiticamente. Introduciamo allora un metodo numerico quadratico di stima degli zeri di una funzione (derivabile), l'algoritmo di Newton-Raphson.

Newton-Raphson nel caso scalare

Sia $l(\vartheta, \vec{x})$ la funzione di log-verosimiglianza (funzione da massimizzare), e sia $S(\vartheta) = l'(\vartheta, \vec{x})$ la funzione score. Noi vogliamo trovare quel $\hat{\vartheta}$ tale per cui $S(\hat{\vartheta}) = 0$. Supponiamo inoltre che S sia derivabile.

Sia $\hat{\vartheta}^{(0)}$ il valore d'inizializzazione dell'algoritmo (di cui parleremo dopo); l'equazione di ricorrenza dell'algoritmo è data da

$$\hat{\vartheta}^{(i+1)} = \hat{\vartheta}^{(i)} - \frac{S(\hat{\vartheta}^{(i)})}{S'(\hat{\vartheta}^{(i)})}$$

processo che si può iterare fino a quando la distanza tra due approssimazioni successive è minore di una certa soglia d'errore fissata.

Valore d'inizializzazione

In generale non esiste una regola che garantisca di scegliere in modo adeguato il valore d'inizializzazione: esso dovrebbe essere scelto il più vicino possibile al valore cercato, ma pure con ciò la convergenza non è garantita, a meno di non porre ulteriori condizioni sulla funzione. (ad esempio, se incontriamo un punto stazionario allora il metodo fallisce in quanto ci troveremmo a dover dividere per zero (e.g. $f(x) = 1 - x^2$, se partiamo da $x^{(0)} = 0$ allora $f'(0) = 0$)). Per approfondire vedasi
http : *//en.wikipedia.org/wiki/Newton's_method#Bad_starting_points*.

Vediamo un paio d'esempi:

Lezione del 27/05, ultima modifica 03/06, Michele Nardin

Capitolo 2

$(X_1, \ldots, X_n)$ campione casuale proveniente da una distribuzione $F_X(x;\theta)$, con $\theta \in \Theta$. Vogliamo verificare un sistema di ipotesi parametriche

$$\begin{cases} H_0 \colon \theta = \theta_0 \\ H_1 \colon \theta = \theta_1 \end{cases}$$

Il sistema di ipotesi è detto *semplice* se esso dipende solamente dal valore di θ, cioè se le ipotesi determinano completamente la distribuzione della statistica.

Abbiamo visto che la funziona di massima verosimiglianza fornisce un "ordinamento" tra i valori assunti da θ, ossia una misura della preferenza di un valore rispetto ad un altro.

$$L(\mathbf{x};\theta) = \prod_{i=1}^{n} f_X(x_i;\theta)\mathbb{I}_S(x_i)$$

In particolare, il rapporto

$$\frac{L(\theta_1;\mathbf{x})}{L(\theta_0;\mathbf{x})}$$

che costituisce una statistica del rapporto di verosimiglianza, può fornire importanti informazioni sulla regione critica del test.

2.1 Teoria dei Test più Potenti ed Uniformemente più Potenti

Useremo la notazione: MP = most powerful e UMP = uniformly most powerful.

Test più potenti per verifica d'ipotesi semplici

Per ora ci concentriamo sui test semplici, ossia i test in cui le ipotesi determinano completamente la distribuzione della statistica sotto la data ipotesi.

Ricordiamo che per la verifica d'ipotesi riguardo ad un parametro ϑ della forma

$$\begin{cases} H_0 : \ \vartheta = \vartheta_0 \\ H_1 : \ \vartheta = \vartheta_1 \end{cases}$$

un dato test fornisce una regione di rifiuto C, (nota: quando ci riferiamo ad un test implicitamente ci riferiamo alla sua regione di rifiuto!) le cui probabilità d'errore sono

$$\begin{cases} \alpha = & P(\text{rifiuto } H_0 | H_0) \\ \beta = & P(\text{non rifiuto } H_0 | H_1) \end{cases}$$

Fissato α, ricordiamo che $1-\beta$ è detto potenza del test. Possiamo sempre pensare β come funzione di ϑ_1:

$$\beta(\vartheta_1) = P(\text{non rifiuto } H_0 | \vartheta = \vartheta_1)$$

Definizione 12 (test MP_α)**.** Fissato $\alpha \in (0,1)$, un test che minimizza $\beta(\vartheta_1)$ è detto test più potente di livello α (MP_α) (per la verifica d'ipotesi semplice contro alternativa semplice).

Il nostro obiettivo sarà ovviamente quello di trovare tale test. Vediamo come i concetti visti fin'ora (soprattutto verosimiglianza e sufficienza) siano legati a quanto cerchiamo tramite il seguente

Lemma 1 (Neyman - Pearson)**.** *Sia $(X_1, \dots, X_n)$ un campione casuale proveniente da una distribuzione di densità $f_{X_i}(\mathbf{x}; \theta)$. Sia inoltre*

$$\begin{cases} H_0 \colon \theta = \theta_0 \\ H_1 \colon \theta = \theta_1 \end{cases}$$

il sistema di ipotesi (entrambe semplici) da verificare. Indicata con $L(\theta; \mathbf{x})$ la funzione di massima verosimiglianza, il test MP_α per la verifica di H_0 ha regione critica (o di rifiuto) data da

$$C = \left\{ \mathbf{x} \in X \colon \frac{L(\theta_1; \mathbf{x})}{L(\theta_0; \mathbf{x})} > A \right\}, \quad A \in \mathbb{R}_+$$

dove A è un valore costante, dipendente da α.

Nota: tale risultato è abbastanza intuitivo: se A è sufficientemente grande, quando si vede che $L(\vartheta_1; \vec{x}) \geq A \cdot L(\vartheta_0; \vec{x})$ è assurdo pensare che ϑ_0 sia una scelta plausibile, in quanto lo è molto di più ϑ_1!

Dimostrazione. Sia C la regione critica del test (di livello α) dato dal lemma, e sia C^* la regione critica di un qualsiasi altro test (sempre di livello α).

Per provare il lemma dobbiamo mostrare quindi che $\beta^* \geq \beta$ (che sono ovviamente le probabilità di errore del 2o tipo delle due regioni in considerazione) (indicheremo nel seguito con $\overline{C}, \overline{C^*}$ i complementari delle regioni C e C^*). Notiamo subito che

$$\overline{C} = \overline{C} \cap (C^* \cup \overline{C^*}) = (\overline{C} \cap C^*) \cup (\overline{C} \cap \overline{C^*})$$

e che

$$\overline{C}^* = \overline{C}^* \cap (C \cup \overline{C}) = (\overline{C}^* \cap C) \cup (\overline{C}^* \cap \overline{C})$$

$$\begin{aligned}
\beta^* - \beta &= P(\vec{x} \in \overline{C}^* | \vartheta = \vartheta_1) - P(\vec{x} \in \overline{C} | \vartheta = \vartheta_1) \\
&= \int_{\vec{x} \in \overline{C}^*} L(\vartheta_1; \vec{x}) d\vec{x} - \int_{\vec{x} \in \overline{C}} L(\vartheta_1; \vec{x}) d\vec{x} \\
&= \int_{\vec{x} \in (\overline{C}^* \cap C) \cup (\overline{C}^* \cap \overline{C})} L(\vartheta_1; \vec{x}) d\vec{x} - \int_{\vec{x} \in (\overline{C} \cap C^*) \cup (\overline{C} \cap \overline{C^*})} L(\vartheta_1; \vec{x}) d\vec{x} \\
&= \int_{\vec{x} \in \overline{C}^* \cap C} L(\vartheta_1; \vec{x}) d\vec{x} - \int_{\vec{x} \in \overline{C} \cap C^*} L(\vartheta_1; \vec{x}) d\vec{x} \\
&\geq A \left[\int_{\vec{x} \in \overline{C}^* \cap C} L(\vartheta_0; \vec{x}) d\vec{x} - \int_{\vec{x} \in \overline{C} \cap C^*} L(\vartheta_0; \vec{x}) d\vec{x} \right] \\
&= A \left[\int_{\vec{x} \in (\overline{C}^* \cap C) \cup (C^* \cap C)} L(\vartheta_0; \vec{x}) d\vec{x} - \int_{\vec{x} \in (\overline{C} \cap C^*) \cup (C \cap C^*)} L(\vartheta_0; \vec{x}) d\vec{x} \right] \\
&= A \left[\int_{\vec{x} \in C} L(\vartheta_0; \vec{x}) d\vec{x} - \int_{\vec{x} \in C^*} L(\vartheta_0; \vec{x}) d\vec{x} \right] \\
&= A(\alpha - \alpha) = 0
\end{aligned}$$

Da cui $\beta^* \geq \beta$ □

Importante: Da nessuna parte abbiamo supposto che ϑ sia uno scalare, e guardando la dimostrazione notiamo che tale ipotesi sarebbe inutile: quindi (come al solito d'altronde) ϑ può benissimo essere un vettore di parametri incogniti!

Test uniformemente più potenti per verifica d'ipotesi semplici contro alternative composte

Vediamo come estendere i risultati della sezione precedente quando le ipotesi alternative sono un pochino più complesse:

A noi interessano soprattutto le ipotesi unilaterali o bilaterali, ossia del tipo

$$\begin{cases} H_0: & \vartheta = \vartheta_0 \\ H_1: & \vartheta > (<) \vartheta_0 \text{ (oppure } \vartheta = \vartheta_1,\ \vartheta_1 > \vartheta_0 (\vartheta_1 < \vartheta_0)) \end{cases}$$

$$\begin{cases} H_0: & \vartheta = \vartheta_0 \\ H_1: & \vartheta \neq \vartheta_0 \text{ (oppure } \vartheta = \vartheta_1,\ \vartheta_1 > \vartheta_0 \vee \vartheta_1 < \vartheta_0) \end{cases}$$

Definizione 13 (UMP_α)**.** Un test per la verifica d'ipotesi semplice (H_0) contro alternativa composta (H_1) viene detto uniformemente più potente di livello α (UMP_α) se esso è MP_α per ogni possibile ipotesi semplice contenuta in H_1 (cioè minimizza $\beta(\vartheta_1)$ per ogni possibile fissato ϑ_1 contemplato in H_1).

Ovviamente nessuno ci assicura che tale test esista: vediamo due esempi, entrambi basati su campione casuale da normale, uno di esistenza (in cui vediamo anche come applicare il lemma) e uno di non esistenza.

Esempio:[*esistenza*] Sia $(X_1, ..., X_n)$ da $N(\mu, \sigma^2)$ (varianza nota). Vogliamo trovare un test UMP_α per le seguenti ipotesi:

$$\begin{cases} H_0: & \mu = \mu_0 \\ H_1: & \mu > \mu_0 \end{cases}$$

Per farlo, considero le ipotesi

$$\begin{cases} H_0: & \mu = \mu_0 \\ H_1: & \mu = \mu_1 \end{cases}$$

per $\mu_1 > \mu_0$ fissato e vado a costruire il test MP_α di Neyman-Pearson associato. Innanzitutto troviamo che

$$\begin{aligned} \frac{L(\mu_1; \vec{x})}{L(\mu_0; \vec{x})} &= exp\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu_1)^2 - (x_i - \mu_0)^2\} \\ &= exp\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n} 2x_i(\mu_0 - \mu_1) + \mu_1^2 - \mu_0^2\} \end{aligned}$$

Ci prepariamo ad applicare il lemma (prendiamo il logaritmo per semplificare i conti): notiamo che

$$-\frac{1}{2\sigma^2}\sum_{i=1}^{n} 2x_i(\mu_0 - \mu_1) + \mu_1^2 - \mu_0^2 \geq ln(A)$$

se e solo se

$$\frac{\sum_{i=1}^{n} x_i}{n} \geq \frac{\sigma^2 ln(A)}{n(\mu_1 - \mu_0)} + \frac{\mu_1 + \mu_0}{2} (=: B)$$

Per rendere operativa la regola di decisione (ossia rifiuto H_0 se $\overline{X_n} \geq B$) occorre specificare il valore di B. Sotto H_0,

$$\overline{X_n} \sim N(\mu_0, \frac{\sigma^2}{n})$$

da cui, fissato un livello di confidenza α, devo risolvere la seguente equazione in B:

$$P(\overline{X_n} \geq B | \mu = \mu_0) = \alpha$$

equivalente a

$$P(\frac{\overline{X_n} - \mu_0}{\sigma/\sqrt{n}} \geq \frac{B - \mu_0}{\sigma/\sqrt{n}} | \mu = \mu_0) = \alpha$$

Ma sotto H_0, $\frac{\overline{X}_n - \mu_0}{\sigma/\sqrt{n}} \sim N(0, 1)$, quindi $\frac{B-\mu_0}{\sigma/\sqrt{n}} = z_\alpha$ (restituito dalle tavole).
Definitivamente, la regione critica sarà

$$C = \{\vec{x} \in X : \overline{x_n} \geq \mu_0 + z_\alpha \frac{\sigma}{\sqrt{n}}\}$$

Abbiamo quindi ritrovato il test che abbiamo usato nel capitolo precedente, senza passare per il concetto di statistica pivot.

Rimane da mostrare che questo test è UMP_α. Ma ciò è già stato fatto, infatti il ragionamento precedente vale quale che sia $\mu_1 (> \mu_0)$ fissato!

Esempio:[*non esistenza*] Sia $(X_1, ..., X_n)$ da $N(\mu, \sigma^2)$ (varianza nota). Vogliamo trovare un test UMP_α per le seguenti ipotesi:

$$\begin{cases} H_0 : & \mu = \mu_0 \\ H_1 : & \mu \neq \mu_0 \end{cases}$$

Seguendo la falsariga di quanto visto nell'esempio precedente, possiamo ovviamente dividere l'ipotesi H_1 nei due seguenti sottocasi:

$$H_1 : \mu = \mu_1,\ \mu_1 > \mu_0 \vee \mu_1 < \mu_0$$

Per il primo caso ($\forall \mu_1 > \mu_0$), come visto sopra, abbiamo la regione critica

$$C = \{\vec{x} \in X : \overline{x_n} \geq \mu_0 + z_\alpha \frac{\sigma}{\sqrt{n}}\}$$

mentre per il secondo caso ($\forall \mu_1 < \mu_0$), in maniera del tutto analoga, avremo

$$C = \{\vec{x} \in X : \overline{x_n} \leq \mu_0 - z_\alpha \frac{\sigma}{\sqrt{n}}\}$$

Notiamo quindi che le due regioni critiche non coincidono su ogni ipotesi semplice contenuta in H_1, e quindi (per definizione) non esiste il test UMP_α per verifica d'ipotesi bilaterali.

Ci poniamo ora il problema di vedere quando esistono i test UMP_α.

Osservazione:[*Legame tra statistiche sufficienti e test di Neyman-Pearson*] Supponiamo di avere un campione casuale da una distribuzione con funzione di densità $f(x; \vartheta)$. Supponiamo inoltre che T_n sia una statistica sufficiente per ϑ. In accordo con il teorema di fattorizzazione, risulta che possiamo scrivere la funzione di verosimiglianza come

$$L(\vartheta; \vec{x}) = h(\vec{x}) g(t_n(\vec{x}), \vartheta)$$

ove h non dipende da ϑ. Quindi, il *rapporto di verosimiglianze (RV)* può essere scritto come

$$RV(\vartheta_1, \vartheta_0; \vec{x}) = \frac{L(\vartheta_1; \vec{x})}{L(\vartheta_0; \vec{x})} = \frac{g(t_n(\vec{x}), \vartheta_1)}{g(t_n(\vec{x}), \vartheta_0)}$$

Segue che, avendo a disposizione una statistica sufficiente per ϑ, tale rapporto dipende da $\vec{x}$ solo attraverso t_n.

Introduciamo il concetto di *monotonia del rapporto di verosimiglianza.*

Definizione 14. Nelle ipotesi dell'osservazione qua sopra, nel caso in cui $\vartheta_1 < \vartheta_0$ ($\vartheta_1 > \vartheta_0$) diciamo che $RV(\vartheta_1, \vartheta_0; \vec{x})$ è monotono se esso è una funzione crescente (decrescente) di t_n.

Intuitivamente, se il rapporto tra le verosimiglianze è monotono, è facile immaginare che il test UMP_α per ipotesi unilaterali esiste sempre!
Ciò è sicuramente vero se ci restringiamo a famiglie esponenziali: consideriamo la famiglia esponenziale nella sua forma più semplice, ossia con funzione di densità della forma

$$f_X(x;\theta) = \exp\{A(\theta)B(x) + C(x) + D(\theta)\}$$

Lemma. *Un campione casuale da una famiglia esponenziale con $A(\vartheta)$ monotona ammette RV monotono.*

Dimostrazione. Infatti,

$$\begin{aligned} RV(\vartheta_1, \vartheta_0; \vec{x}) &= \frac{\exp\{A(\vartheta_1)\sum_i B(x_i) + \sum_i C(x_i) + nD(\vartheta_1)\}}{\exp\{A(\vartheta_0)\sum_i B(x_i) + \sum_i C(x_i) + nD(\vartheta_0)\}} \\ &= \exp\{[A(\vartheta_1) - A(\vartheta_0)] \cdot \sum_i B(x_i) + n[D(\theta_1) - D(\theta_0)]\} \end{aligned}$$

Essendo A monotona, si nota subito che se $\vartheta_1 > \vartheta_0$ allora $A(\vartheta_1) - A(\vartheta_0) \geq 0$ e quindi RV è una funzione crescente rispetto a $\sum_i B(x_i)$ □

Da quanto visto sopra, segue immediatamente che, testando le ipotesi

$$\begin{cases} H_0 : & \vartheta = \vartheta_0 \\ H_1 : & \vartheta < \vartheta_0 \end{cases}$$

preso un qualunque $\vartheta_1 < \vartheta_0$, la condizione

$$RV(\vartheta_1, \vartheta_0; \vec{x}) \leq k$$

è equivalente a

$$\sum_{i=1}^{n} B(x_i) \leq c$$

per una costante c ricavabile facilmente dall'equazione sopra, e valido per ogni $\vartheta_1 < \vartheta_0$. Questo fornisce il tesi UMP_α per la famiglia esponenziale (in questo caso vista nella forma semplice) per test unilaterali. Ovviamente, nel caso in cui $H_1 : \vartheta > \vartheta_0$, la regione critica sarà fornita da $\sum_{i=1}^{n} B(x_i) \geq c$.

2.1.1 Rapporto di massime verosimiglianze

Introduciamo un test valido in generale, con il defetto di esser senza garanzie (se non asintotiche) di ottimalità ma con il pregio di non aver bisogno di particolari forme d'ipotesi.

Definizione 15. Sia $(X_1, ..., X_n)$ un campione casuale da una distribuzione avente funzione di ripartizione $F(x, \vartheta)$, con $\vartheta \in \Theta$. Supponiamo di avere un sistema di ipotesi

$$\begin{cases} H_0 : & \vartheta \in \Theta_0 \\ H_1 : & \vartheta \in \Theta \setminus \Theta_0 \end{cases}$$

Definiamo la funzione Rapporto di Verosimiglianza generalizzato (RV generalizzato) come la funzione $\lambda : \mathcal{X} \in \mathbb{R}^n \to [0, 1]$ definita da

$$\lambda(\vec{x}) = \frac{\max_{\vartheta \in \Theta_0} L(\vartheta, \vec{x})}{\max_{\vartheta \in \Theta} L(\vartheta, \vec{x})}$$

λ sta in $[0, 1]$ poiché, essendo L una funzione di densità, $\lambda \geq 0$, ed invece $\lambda \leq 1$ essendo $\Theta_0 \subset \Theta$.

Fissato un $A \in (0, 1)$, una regione di rifiuto del test è data da

$$C := \{\vec{x} \in \mathcal{X} : \lambda(\vec{x}) < A\}$$

Ovviamente A sarà scelto in modo che

$$P(\lambda(\vec{x}) < A | H_0) = \alpha$$

con α probabilità (fissata a priori) di commettere un errore del primo tipo.

La potenza di questo metodo è data dal seguente

Teorema 24. *Sotto H_0 la statistica (test di Wald) $W = -2ln[\lambda(\vec{x})]$, converge in distribuzione (all'aumentare di n) a χ^2_ν, ove ν è la differenza tra i parametri da stimare sotto H_1 rispetto ad H_0.*

Quindi, per campioni numerosi, è possibile ricavare agevolmente il parametro A dalle tavole della χ^2.

Esempio: Sia $(X_1, ..., X_n)$ un campione casuale da Poisson con parametro θ, e quindi

$$f(x, \theta) = P(X_i = x) = \frac{e^{-\theta}\theta^x}{x!}\mathbb{I}_{(0,1,2,...)}(x)$$

e $\theta > 0$.

$$\begin{cases} H_0 : & \theta = \theta_0 \\ H_1 : & \theta \neq \theta_0 \quad [\theta \in \mathbb{R}^+ \setminus \{\theta_0\}] \end{cases}$$

Calcoliamo inoltre la funzione di verosimiglianza:

$$L(\theta, \vec{x}) = \frac{e^{-n\theta}\theta^{\sum x_i}}{\prod_{i=1}^n x_i!}$$

troviamo che, posto $\hat{\theta} = \overline{X}_n$ lo stimatore di massima verosimiglianza

$$\lambda(\vec{x}) = \frac{\max_{\theta \in \Theta_0} L(\theta, \vec{x})}{\max_{\theta \in \Theta} L(\theta, \vec{x})} = \frac{L(\theta_0, \vec{x})}{L(\hat{\theta}, \vec{x})}$$

(notare che $\max_{\theta\in\Theta} L(\theta,\vec{x}) = L(\hat{\theta},\vec{x})$ proprio per definizione di stimatore di massima verosimiglianza!)

$$\lambda(\vec{x}) = \frac{e^{-n\theta_0}\theta_0^{\sum x_i}}{e^{-n\hat{\theta}}\hat{\theta}^{\sum x_i}} = \left(\frac{\theta_0}{\hat{\theta}}\right)^{\sum x_i} e^{n(\hat{\theta}-\theta_0)}$$

Usiamo la statistica test di Wald: $W = -2ln(\lambda(\vec{x})) = -2\left[\sum x_i ln\left(\frac{\theta_0}{\hat{\theta}}\right) + n(\hat{\theta}-\theta_0)\right]$ la quale, sotto H_0, $W \xrightarrow{D} \chi^2$. La regione critica che cerchiamo è data da

$$C := \{\vec{x}\in\mathcal{X} : \lambda(\vec{x}) < A\}$$

Vale $\lambda(\vec{x}) < A \Leftrightarrow -2ln[\lambda(\vec{x})] > -2ln(A)$. Fissato α, se n è abbastanza grande possiamo approssimare

$$P(\lambda(\vec{x}) < A \;\Big|\; H_0) = P(-2ln[\lambda(\vec{x})] > -ln(A^2) \;\Big|\; H_0) \cong P(\chi_1^2 > -ln(A^2) \;\Big|\; H_0) = \alpha$$

sse

$$-ln(A^2) = \chi^2_{1;\alpha/2}$$

da cui

$$C = \{\vec{x}\in\mathcal{X} : -2\left[\sum x_i ln\left(\frac{\theta_0}{\hat{\theta}}\right) + n(\hat{\theta}-\theta_0)\right] > \chi^2_{1;\alpha/2}\}$$