Modelli e Metodi Dell'inferenza Statistica

Dalle lezioni della Prof. Anna Maria Paganoni

per il corso di Ingegneria Matematica

Appunti di Simone Paloschi

Politecnico di Milano

A.A. 2022/2023

Indice

1 Ripasso dei fondamentali

Lezione 1 (20/02/2023)

- Delle variabili aleatorie (va) si dicono indipendenti e identicamente distribuite (iid) se sono a 2 a 2 indipendenti e hanno tutte la stessa legge

- Dati dicotomici sono dati che hanno un successo e un insuccesso (due risultati)

Lezione 2 (21/02/2023)

- Media e varianza campionarie: $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \quad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2$

- q_α quantile di ordine α di una va X continua è t.c. $\mathbb{P}(X \leq q_\alpha) = \alpha$

Oss. Posti t_α e z_α i quantili di una t-student e una normale vale sempre $t_\alpha(m) > z_\alpha \ \ \forall m$ ordine della t

- Leggi condizionate: $\mathbb{E}[X|Y] = g(Y)$ è una va che è funzione di Y e vale $\mathbb{E}[\mathbb{E}[X|Y]] = \mathbb{E}[X]$

Inoltre vale la decomposizione della varianza $\mathrm{Var}[X] = \mathrm{Var}[\mathbb{E}[X|Y]] + \mathbb{E}[\mathrm{Var}[X|Y]]$

- Funzione generatrice dei momenti $m_X(t) = \mathbb{E}\left[e^{tX}\right]$

Oss. Calcolando il valore della sua derivata k-esima in 0 ottengo il momento di ordine k di X

Convergenze:

$X_n \xrightarrow{qc} X$

$X_n \xrightarrow{P} X \quad lim_{\varepsilon\to 0}\mathbb{P}((|X_n - X| < \varepsilon)) \to 1$

$X_n \xrightarrow{\mathcal{L}} X \quad F_{X_n}(t) \xrightarrow[n\to\infty]{} F_X(t)\ \forall t$ di continuità di F_X

$qc \implies P \implies \mathcal{L}$ al contrario c'è implicazione da legge a probabilità se converge a una costante

Teo di Slutsky: Se $X_n \xrightarrow{\mathcal{L}} X \ \ Y_n \xrightarrow{P} K$, allora $X_n + Y_n \xrightarrow{\mathcal{L}} X + K$ e $X_n \cdot Y_n \xrightarrow{\mathcal{L}} X \cdot K$

Distribuzioni:

- Gaussiana: $Z \sim \mathcal{N}\left(\mu, \sigma^2\right)$ $\mathbb{E}\left[X_i\right] = \mu$ e $\mathrm{Var}\left[X_i\right] = \sigma^2$ $f_Z(x) = \dfrac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\dfrac{(x-\mu)^2}{2\sigma^2}\right\}$

- Campione gaussiano: $X_1, ..., X_n$ con $X_i \sim \mathcal{N}(\mu, \sigma^2)$ valgono:
a) $\overline{X}_n \sim \mathcal{N}\left(\mu, \dfrac{\sigma^2}{n}\right)$ *b)* $\overline{X}_n \perp\!\!\!\perp S^2$ *c)* $\dfrac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$

- Gamma: $X \sim \Gamma(\alpha, \beta)$ $f_X(x) = \dfrac{\beta^\alpha x^{\alpha-1} e^{-\beta x}}{\Gamma(\alpha)} \mathbb{1}_{[0,+\infty)}(x)$ valgono:
a) $\Gamma\left(\dfrac{k}{2}, \dfrac{1}{2}\right) \sim \chi^2(k)$ *b)* $Y = hX$ con $h > 0 \implies Y \sim \Gamma\left(\alpha, \dfrac{\beta}{n}\right)$

- t-student: $T \sim t(m)$ $f_T(t) = \dfrac{\Gamma\left(\dfrac{m+1}{2}\right)}{\Gamma\left(\dfrac{m}{2}\right)\sqrt{\pi m}} \cdot \dfrac{1}{\left(1 + \frac{t^2}{m}\right)^{\frac{m+1}{2}}}$ valgono:
a) $\dfrac{\overline{X}_n - \mu}{\frac{S}{\sqrt{n}}} \sim t(n-1)$ *b)* $f_T(t) \xrightarrow{m\to\infty} f_Z(z) = \dfrac{1}{\sqrt{2\pi}}\ exp\{-\dfrac{t^2}{2}\}$
c) $T \sim t(m)$ è simmetrica rispetto allo 0 e all'infinito va a zero come t^{m+1}, quindi $\mathbb{E}\left[T^k\right] < \infty \Leftrightarrow k < m$

- Distribuzione di Fischer: $U \sim \chi^2(n)$ $V \sim \chi^2(m)$ $U \perp\!\!\!\perp V$
$X = \dfrac{\frac{U}{n}}{\frac{V}{m}} \sim F(n,m)$ $f_X(x) = \dfrac{\Gamma\left(\dfrac{n+m}{2}\right)}{\Gamma\left(\dfrac{n}{2}\right)\Gamma\left(\dfrac{m}{2}\right)} m^{\frac{m}{2}} n^{\frac{n}{2}} \dfrac{x^{\frac{n}{2}-1}}{(nx+m)^{\frac{n+m}{2}}} \mathbb{1}_{[0,+\infty)}(x)$

Teo: a) $X \sim F(n,m) \implies \dfrac{1}{X} \sim F(n,m)$ b) $X \sim t(m)$ $X^2 \sim F(1,m)$

- T.C.L. media campionaria quando ho un campione qualsiasi (anche non gaussiano):
$\sqrt{n}(\overline{X}_n - \mu) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \sigma^2)$

- Dato un campione $X_1, ..., X_n$ e poste le va $X_{(1)} = \underset{i=1..n}{Min}\ X_i$ $X_{(n)} = \underset{i=1..n}{Max}\ X_i$ valgono:
$F_{X_{(n)}}(t) = \mathbb{P}\left(X_{(n)} \le t\right) = \mathbb{P}\left(X_i \le t\ \forall i\right) = \mathbb{P}\left(\bigcap_{i=1}^n\ X_i \le t\right) = \{\text{indip}\} = \prod_{i=1}^n \mathbb{P}\left(X_i \le t\right) = \{\text{i.i.d}\} = \left(F_X(t)\right)^n$
$F_{X_{(1)}}(t) = \mathbb{P}\left(X_{(1)} \le t\right) = 1 - \mathbb{P}\left(X_{(1)} > t\right) = \{\text{ se iid }\} = ... = 1 - \left[\mathbb{P}\left(X > t\right)\right]^n = 1 - \left[1 - F_X(t)\right]^n$

Oss. Se $X_i \overset{iid}{\sim} \mathcal{E}(\lambda) \implies X_{(1)} \sim \mathcal{E}(\lambda n)$ perché $F_{X_i}(t) = 1 - e^{-\lambda t} \implies F_{X_{(1)}} = 1 - e^{-\lambda n t}$

2 Statistica inferenziale parametrica

Esempio: Data una moneta (v.a. X Bernulliana), l'obbiettivo della statistica è quello di trovare il parametro $p = \mathbb{P}(X = 1)$. L'unico modo che ho è quello di lanciare la moneta

Evento: $X_1, ..., X_n \overset{\omega}{\rightsquigarrow} x_1, ..., x_n$ dove $x_i = X_i(\omega)$

Definizione.

Un **Campione casuale** è un insieme di v.a. i.i.d. $X_1, ..., X_n$, dove n è l'ampiezza del campione

Definizione.

Dato un campione casuale $X_1, ..., X_n$ la legge $\mathcal{L}(X_i)$ è **parametrica** se è nota a meno di un numero finito di parametri

Oss. Se non fosse parametrica, dovremmo cercare la legge di X_i tra tutte le funzioni di ripartizione, ovvero in un insieme infinito dimensionale. Sarebbe un problema troppo complicato, se invece cerco tra le leggi parametriche restringo il problema, perché $F_X(t)$ dipende da $\vec{\theta} = \theta_1, ..., \theta_k$ con $k < \infty$

Esempi: a) $F_X(t) \sim Be(p)$ ho $\theta_1 = p$ quindi $k = 1$ b) $F \sim \mathcal{N}(\mu, \sigma^2)$ $k = 2$
Per trovare la funzione mi basta trovare questi parametri

Definizione.

Un **modello statistico** $\left(\mathbb{R}^n;\ \mathcal{B}(\mathbb{R}^n);\ \mathbb{P}_{\vec{\theta}}\right)$ è lo spazio dove assumono i valori i miei campioni casuali con $\vec{\theta} \in \Theta$ spazio dei parametri

Definizione.

$Y = T(X_1, ..., X_n)$ è una **statistica**, cioè una qualsiasi funzione del campione

Esempi: $\sum X_i$ $\prod X_i$ $X_{(1)}$ S^2 sono statistiche
$\dfrac{(n-1)S^2}{\sigma^2}$ non è una statistica perché dipende da σ che è un parametro che devo trovare

La legge di Y è detta legge campionaria
Minimo, massimo... $X_{(1)}...X_{(k)}...X_{(n)}$ sono dette statistiche d'ordine
Oss. Ogni statistica è una riduzione dei dati e ci da informazioni sul campione, per esempio il minimo e il massimo prendono informazioni da $\mathbb{R}^n$ e restituiscono informazioni utili in $\mathbb{R}$

Lezione 3 (22/02/2023)

La realizzazione del campione su ω è $t = Y(\omega) = T(X_1(\omega), ..., X_n(\omega))$

Un'inferenza è un processo di ricerca dei parametri $\vec{\theta}$

Un'inferenza di $T(X)$ su θ è il valore di θ che trovo per un valore $x = X(\omega)$

Teorema: Principio di sufficienza.

Una statistica $T(\vec{X})$ è sufficiente per θ se ogni inferenza su θ dipende dal campione $\vec{X}$ solo tramite $T(\vec{X})$

Prop. Se $\vec{x}, \vec{y}$ sono due realizzazioni diverse del campione t.c. $T(\vec{x}) = T(\vec{y})$ e se vale il principio di sufficienza, allora l'inferenza su θ sarà identica sia che osservi $\vec{x}$ o $\vec{y}$

Definizione.

$T(\vec{x})$ è una statistica **sufficiente** per θ se la distribuzione di $\vec{X} = (X_1, ..., X_n)$ dato $t = T(\vec{X})$ non dipende da θ per qualunque valore di t

Oss. La def equivale a dire $\mathcal{L}\left(\vec{X} \mid T(\vec{X}) = t\right)$ non dipende da $\theta \quad \forall t$

Ed equivale a chiedere che valga il principio di sufficienza

Esempio: $X_1, X_2 \sim Be(\theta)$ allora $T = X_1 + X_2$ è una stat suff

Per verificarlo devo calcolare la legge condizionata $\mathcal{L}\left(X_1, X_2 \mid T(\vec{X} = t)\right) \quad \forall t$ cioè $t = 0, 1, 2$

$\mathbb{P}(X_1 = 0, X_2 = 0 \mid T = 0) = 1 \quad \mathbb{P}(X_1 = 1, X_2 = 1 \mid T = 2) = 1$

$\mathbb{P}(X_1 = 1, X_2 = 0 \mid T = 1) = \mathbb{P}(X_1 = 0, X_2 = 1 \mid T = 1) = \dfrac{1}{2}$

Quindi $\forall t$ la legge condizionata non dipende da θ (ovvero non compare θ nelle probabilità condizionate)

Teorema: Criterio di fattorizzazione.

Sia $f(\vec{x}, \theta)$ la densità di probabilità congiunta di $\vec{X}$, una statistica $T(\vec{X})$ è sufficiente per θ se e solo se esistono due funzioni $g(t, \theta)$ e $h(\vec{x})$ t.c. $\quad f(\vec{x}, \theta) = h(\vec{x})\ g(T(\vec{x}), \theta) \quad \forall \vec{x}\ \forall \theta$

Esempio: $X_1, ..., X_n \sim Be(p) \quad T(\vec{X}) = \sum_{i=1}^{n} X_i$ è suff per p verifichiamolo con il criterio di fattorizzazione

$f(\vec{x}, p) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i}\ \mathbb{1}_{\{0,1\}}(x_i) = p^{\sum x_i}(1-p)^{n-\sum x_i} \prod_{i=1}^{n} \mathbb{1}_{\{0,1\}}(x_i) = g(\sum x_i, p)\ h(\vec{x})$

Oss. La densità la scrivo così perché equivale a moltiplicare per p quando $x_i = 1$ e per (1-p) quando $x_i = 0$

Dim.

Criterio di fattorizzazione (nel caso discreto, senza perdita di generalità w.l.g.)

1) Sia $T(\vec{X})$ stat sufficiente per θ, allora

$$f(\vec{x},\theta) = \mathbb{P}_\theta\left(\vec{X} = \vec{x} \mid T(\vec{X}) = T(\vec{x})\right) \cdot \mathbb{P}_\theta\left(T(\vec{X}) = T(\vec{x})\right) = h(\vec{x}) \cdot g(T(\vec{x},\theta))$$

Perchè il primo fattore non dipende da θ essendo T sufficiente

2)* Assumiamo che valga la fattorizzazione

Sia $q(t,\theta)$ la densità di prob di $T(\vec{X})$

Sia $A_{T(\vec{x})} = \{\vec{y} : T(\vec{y}) = T(\vec{x})\} =$ l'insieme dei vettori che hanno la stessa controimmagine di $\vec{x}$

$$\mathcal{L}\left(\vec{X} \mid T(\vec{x})\right) = \frac{f(\vec{x},\theta)}{q(T(\vec{x},\theta))} = \frac{g(T(\vec{x}),\theta)h(\vec{x})}{\sum\limits_{\vec{y}\in A_{T(\vec{x})}} f(\vec{y},\theta)} = \frac{g(T(\vec{x},\theta),\theta)h(\vec{x})}{\sum\limits_{\vec{y}\in A_{T(\vec{x})}} g(T(\vec{y}),\theta)h(\vec{y})} =$$

$$= \frac{g(T(\vec{x}),\theta)h(\vec{x})}{g(T(\vec{x}),\theta)\sum\limits_{\vec{y}\in A_{T(\vec{x})}} h(\vec{y})} = \frac{h(\vec{x})}{\sum\limits_{\vec{y}\in A_{T(\vec{x})}} h(\vec{y})} \text{ che non dipende da } \theta \implies \text{T è suff}$$

□

Esempio: $X_1,...,X_n \sim \mathcal{U}([0,\theta])$ è l'uniforme

$f_{X_i}(x_i,\theta) = \frac{1}{\theta}\mathbb{1}_{[0,\theta]}(x_i) \quad f(\vec{x},\theta) = \prod_{i=1}^n f_{X_i}(x_i,\theta) = \frac{1}{\theta^n}\prod_{i=1}^n \mathbb{1}_{[0,\theta]}(x_i)$

La produttoria di indicatrici è un indicatrice, che vale 1 se $\forall i\ 0 \le x_i \le \theta$, quindi posso lavorare con il massimo

$f(\vec{x},\theta) = \frac{1}{\theta^n}\mathbb{1}_{[0,\theta]}(X_{(n)}) \cdot 1 = g(X_{(n)},\theta)\cdot h(\vec{x}) \implies T(\vec{x}) = X_{(n)} = \underset{i=1...n}{Max} X_i$ è sufficiente

Esempio: $X_1,...,X_n \sim \mathcal{N}(\mu,\sigma^2) \quad \vec{\theta} = (\mu,\sigma^2) \ \ \Theta = \mathbb{R}\times\mathbb{R}^+$

$$f(\vec{x},\vec{\theta}) = \frac{1}{(\sqrt{2\pi\sigma^2})^n}\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i-\mu)^2\right\} = \frac{1}{(\sqrt{2\pi\sigma^2})^n}\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^n x_i^2 - \frac{n\mu^2}{2\sigma^2} + \frac{\mu}{\sigma}\sum_{i=1}^n x_i\right\} =$$

$$= g(\sum X_i, \sum X_i^2, \vec{\theta}) = g(T_1(\vec{(}X)), T_2(\vec{X}), \theta)$$

$\vec{T}(\vec{X}) = (\sum X_i, \sum X_i^2)$ è una stat suff bivariata per θ (bivariata vuol dire che sta in $\mathbb{R}^2$)

Oss. Analogamente posso provare che $(\vec{X}_n, S^2)$ è una stat suff bivariata per θ

Teorema.

Sia r una funzione biunivoca e $T(\vec{X})$ una stat suff, allora $T^*(\vec{X}) = r(T(\vec{X}))$ è suff

Dim.

$$f(\vec{x},\theta) = g(T(\vec{x}),\theta)h(\vec{x}) = g(r^{-1}(T^*(\vec{x})),\theta)h(\vec{x}) = g^*(T^*(\vec{x}),\theta)h(\vec{x})$$

□

Lezione 4 (24/02/2023)

Definizione.

Una va X appartiene alla **famiglia esponenziale**: $X \in EF$ se $f(x,\vec{\theta}) = h(x)c(\vec{\theta})\, exp\left\{\sum_{j=1}^{k} t_j(x)w_j(\vec{\theta})\right\}$

Oss. Da $f(\vec{x},\vec{\theta}) = \prod_{i=1}^{n} h(x_i)\; c(\vec{\theta})^n\, exp\left\{\sum_{i=1}^{n}\sum_{j=1}^{k} t_j(x_i)w_j(\vec{\theta})\right\}$ ottengo subito che
$\vec{T}(\vec{X}) = \left(\sum_{i=1}^{n} t_1(X_i), ..., \sum_{i=1}^{n} t_k(X_i)\right)$ è suff per $\vec{\theta}$

Esempio: $X \sim Be(p)$ $f(x,p) = p^x(1-p)^{1-x}\mathbb{1}_{\{0,1\}}(x) = \mathbb{1}_{\{0,1\}}(x)\cdot(1-p)\, exp\left\{x\, log(\frac{p}{1-p})\right\} = h(x)\cdot c(p)\cdot e^{\cdots}$
In questo caso abbiamo $k=1$ $t_1(x) = X$ $w_1(p) = log(\frac{p}{1-p})$ $\implies$ $T(\vec{X}) = \sum_{i=1}^{n} X_i$ è suff per p

Esempio: $X \sim \mathcal{P}(\lambda)$ $f(x,\lambda) = \frac{e^{-\lambda}\lambda^x}{x!}\mathbb{1}_{\mathbb{N}}(x) = \frac{\mathbb{1}_{\mathbb{N}}(x)}{x!}e^{-\lambda}\, exp\{x\, log(\lambda)\}$ $\implies$ $T(\vec{X}) = \sum X_i$ è suff per λ

Esempio: Facile verificare che anche $X \sim \mathcal{E}(\lambda) \in EF$ e $X \sim \Gamma(n,\lambda) \in EF$

Esempio: $X \sim \mathcal{N}(\mu,\sigma^2)$ $f(x,\vec{\theta}) = \frac{1}{\sqrt{2\pi\sigma^2}}\, exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\} = \frac{1}{\sqrt{2\pi\sigma^2}}\, exp\left\{-\frac{\mu^2}{2\sigma^2}\right\}\left\{-\frac{x^2}{2\sigma^2} + \frac{\mu}{\sigma^2}x\right\}$
In questo caso abbiamo $k=2$ $t_1(x) = x^2$ $t_2(x) = x$ $\implies$ $\vec{T}(\vec{X}) = (\sum X_i, \sum X_i^2)$ è suff
Oss. Avevo ottenuto lo stesso risultato con la fattorizzazione, ma era molto più laborioso

Esempio: $X \sim \mathcal{U}_{[0,\theta]}(x)$ $f(x,\theta) = \frac{1}{\theta}\mathbb{1}_{[0,\theta]}(x)$ non sta nella famiglia esponenziale, perché $c(\theta) = \frac{1}{\theta}$ ma l'indicatrice, in nessun modo, si può scrivere come un esponenziale

Definizione.

Una statistica sufficiente $T(\vec{X})$ è detta **sufficiente e minimale** se per ogni altra stat suff $T'(\vec{X})$, $T(\vec{X})$ è una funzione di $T'(\vec{X})$, ovvero $\forall \vec{x}, \vec{y}\ tc\ T'(\vec{x}) = T'(\vec{y}) \implies T(\vec{x}) = T(\vec{y})$

Oss. Tutto il campione è sufficiente, però non è minimale. Infatti ha più informazioni della media campionaria e per questo non si può scrivere come funzione della media, che è una statistica sufficiente

Teorema: Lemma di Scheffè 1.

Sia $f(\vec{x}, \theta)$ la densità congiunta di $\vec{X}$. Se esiste una funzione $T(\vec{X})\ tc\ \forall \vec{x}, \vec{y}$ il quoziente $\dfrac{f(\vec{x}, \theta)}{f(\vec{y}, \theta)}$ è costante rispetto a θ se e solo se $T(\vec{x}) = T(\vec{y})$, allora $T(\vec{X})$ è stat suff e minimale per θ

Esempio: $X_1, ..., X_n \sim \mathcal{U}_{[0,\theta]}\ \ f(\vec{x}, \theta) = \dfrac{1}{\theta_n} \prod_{i=1}^{n} \mathbb{1}_{[0,\theta]}(x_i) = \dfrac{1}{\theta^n} \mathbb{1}_{[0,\theta]}(x_{(n)})$
$\dfrac{f(\vec{x}, \theta)}{f(\vec{y}, \theta)} = \dfrac{\mathbb{1}_{[0,\theta]}(x_{(n)})}{\mathbb{1}_{[0,\theta]}(y_{(n)})}$, questo quoziente non dipende da θ se e solo se $T(\vec{x}) = x_{(n)} = T(\vec{y}) = y_{(n)}$, allora $X_{(n)}$ è suff e minimale

Esempio: $X_1, ..., X_n \sim \mathcal{N}(\mu, \sigma^2)$

$$\frac{f(\vec{x}, \vec{\theta})}{f(\vec{y}, \vec{\theta})} = \frac{\dfrac{1}{\left(\sqrt{2\pi\sigma^2}\right)^n} exp\left\{-\dfrac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\right\}}{\dfrac{1}{\left(\sqrt{2\pi\sigma^2}\right)^n} exp\left\{-\dfrac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu)^2\right\}} = exp\left\{-\frac{\mu^2}{2\sigma^2} - \frac{\sum x_i^2}{2\sigma^2} + \frac{\mu}{\sigma^2}\sum x_i + \frac{\mu^2}{2\sigma^2} + \frac{\sum y_i^2}{2\sigma^2} + \frac{\mu}{\sigma^2}\sum y_i\right\} =$$

$$= exp\left\{-\frac{\sum x_i^2}{2\sigma^2} + \frac{\mu}{\sigma^2}\sum x_i + \frac{\sum y_i^2}{2\sigma^2} + \frac{\mu}{\sigma^2}\sum y_i\right\} \quad \text{non dipende da } (\mu, \sigma^2) \iff \begin{cases} \sum x_i^2 = \sum y_i^2 \\ \sum x_i = \sum y_i \end{cases}$$

$\implies \vec{T}(\vec{X}) = (\sum X_i, \sum X_i^2)$ è stat suff e minimale

Teorema.

Ogni funzione biunivoca di una statistica suff e minimale è anch'essa suff e minimale

Esempio: $\vec{T}(\vec{X}) = (\sum X_i, \sum X_i^2)$ è suff e minimale $\implies (\overline{X}_n, S^2)$ è suff e minimale

Definizione.

Sia $T(\vec{X})$ una statistica e sia $h(t, \vec{\theta})$ la famiglia di densità di T, questa famiglia si dice **completa** (e quindi dirò che $T(\vec{X})$ è completa) se $\mathbb{E}_\theta[g(T)] = 0\ \forall \theta \implies \mathbb{P}(g(T) = 0) = 1$

Esempio: $X_1, ..., X_n \sim Be(p)$ $T(\vec{X}) = \sum X_i \sim Bi(n,p)$

Sia g una funzione di T tc $\mathbb{E}_p[g(T)] = 0 \ \forall p \implies \sum_{k=0}^{n} g(k)\binom{n}{k}p^k(1-p)^{n-k} = \sum_{k=0}^{n} g(k)\binom{n}{k}(\frac{p}{1-p})^k(1-p)^n = 0 \ \forall p$

Posto $s = \left(\frac{p}{1-p}\right)^k$ quindi $\sum_{k=0}^{n} g(k)\binom{n}{k}s^k = 0 \ \forall s$, questo è un polinomio di grado n in s identicamente nullo

$\implies g(k)\binom{n}{k} = 0 \ \forall k = 0...n \implies g(k) = 0 \ \forall k \implies \mathbb{P}(g(T) = 0) = 1 \implies T = \sum X_i$ è completa

Esempio: $X_1, ..., X_n \sim \mathcal{U}_{[0,\theta]}$ $X_{(n)}$ è suff e min, verifico completezza

Mi serve la densità del massimo, so che $F_{X_{(n)}}(t) = (F_X(t))^n$

Essendo $f_X(t) = \frac{1}{\theta}\mathbb{1}_{[0,\theta]}$ avrò $F_X(t) = \frac{t}{\theta}\mathbb{1}_{[0,\theta]} + \mathbb{1}_{[0,+\infty]}(t)$

$F_{X_{(n)}}(t) = \left(\frac{t}{\theta}\right)^n \mathbb{1}_{[0,\theta]} + \mathbb{1}_{[0,+\infty]}(t)$ ottengo $f_{X_{(n)}}(t,\theta) = n\frac{t^{n-1}}{\theta^n}\mathbb{1}_{[0,\theta]}(t) = h(t,\theta)$

$\forall\theta \ 0 = \mathbb{E}_\theta[g(X_{(n)})] = \int_{\mathbb{R}} g(t)h(t,\theta)\,dt \implies \frac{d}{d\theta}\mathbb{E}_\theta[g(X_{(n)})] = 0 = \frac{d}{d\theta}\int_0^\theta g(t)\frac{n}{\theta^n}t^{(n-1)}\,dt = \frac{d}{d\theta}\left[\frac{1}{\theta^n}\int_0^\theta g(t)nt^{(n-1)}\,dt\right]$

$\implies 0 = \frac{1}{\theta^n}g(\theta)n\theta^{(n-1)} + \frac{d}{d\theta}\left(\frac{1}{\theta^n}\right)\cdot\mathbb{E}[g(X_{(n)})]\cdot\theta^n$ però il valore atteso è nullo

$\implies \forall\theta \ \frac{ng(\theta)}{\theta} = 0 \implies g(\theta) = 0 \ \forall\theta \implies \mathbb{P}\big(g(X_{(n)}) = 0\big) = 1$

Lezione 5 (28/02/2023)

> **Teorema: Bahadur.**
>
> Una statistica suff e completa è anche minimale

Oss. Questo è utile perché verificare la completezza è spesso più facile della minimalità

Oss. Abbiamo visto che dato $X_1, ..., X_n$ con $\mathcal{L}(X_i) \in EF$ ovvero $f(x,\vec{\theta}) = h(x)c(\vec{\theta})\,exp\left\{\sum_{j=1}^{k} t_j(x)w_j(\vec{\theta})\right\}$

Allora $\vec{T}(\vec{X}) = \left(\sum_{i=1}^{n} t_1(X_i), ..., \sum_{i=1}^{n} t_k(X_i)\right)$ è suff per $\vec{\theta}$

> **Teorema.**
>
> $\vec{T}(\vec{X})$ è completa se $\{w_1(\vec{\theta}), ..., w_k(\vec{\theta})\}$ mappa Θ in un insieme che contiene almeno un aperto di $\mathbb{R}^k$

Esempio: $X_i \sim \mathcal{P}(\lambda)$ $f(x,\lambda) = \frac{\mathbb{1}_{\mathbb{N}}(x)}{x!}e^{-\lambda}\,exp\{x\,log(\lambda)\} \implies t_1(x) = x$

$w_1(\lambda) = log(\lambda) : \mathbb{R}^+ \to \mathbb{R}$ che contiene aperti $\implies$ è completa (e per Bahadur anche minimale)

Esempio: $X_1, ..., X_n \sim \mathcal{N}(\mu,\sigma^2)$

$$f(x,\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\,exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\} = \frac{1}{\sqrt{2\pi\sigma^2}}\,exp\left\{-\frac{\mu^2}{2\sigma^2}\right\}\,exp\left\{-\frac{x^2}{2\sigma^2} + \frac{\mu}{\sigma^2}x\right\}$$

$$t_1(x) = x^2 \quad t_2(x) = x \quad w_1(\vec{\theta}) = -\frac{1}{2\sigma^2} \quad w_2(\vec{\theta}) = \frac{\mu}{\sigma^2} \qquad \vec{\theta} = (\mu, \sigma^2) \in \Theta = \mathbb{R} \times \mathbb{R}^+$$

Quindi $(w_1, w_2) : \Theta \to \mathbb{R}^- \times \mathbb{R}$ che contiene aperti $\implies (\sum X, \sum X_i^2)$ è suff minimale completa

Esempio: $X_1, ..., X_n \sim \mathcal{N}(\mu, \mu^2)$ ottengo analogamente $w_1 = -\frac{1}{2\mu^2} \quad w_2 = \frac{1}{\mu}$ queste mappano una parabola, che non contiene un aperto di $\mathbb{R}^2$, quindi non possiamo concludere che sia completa

Oss. In questo esempio posso vedere che il campione dipende solo da un parametro e concludere subito che non posso applicare il teorema

Definizione.

Una statistica $S(\vec{X})$ è detta **ancillare** se ha la distribuzione che non dipende da θ

Esempio: $X_1, ..., X_n \sim \mathcal{U}([\theta, \theta + 1])$

Considero il $Range = X_{(n)} - X_{(1)} \sim Beta(n - 1, 2)$ è ancillare

Prop. "In molte situazioni" Se S è una statistica ancillare e T è stat suff, minimale e completa, allora $S \perp\!\!\!\perp T$

3 Stima puntuale dei parametri

Definizione.

Uno **stimatore** è un oggetto che produce stime di $\vec{\theta}$

Il nostro compito sarà quello di:

- Costruire gli stimatori
- Definire metodi per valutare gli stimatori

Definizione.

Uno stimatore puntuale per θ è una qualunque funzione $W(X_1, ..., X_n)$ del campione

Oss. Ovvero uno stimatore è una statistica che usiamo per stimare θ

Oss. Stimatore $\neq$ stima puntuale, che è la "realizzazione" dello stimatore: $\mathrm{w} = W(\vec{X}(\omega)) = W(x_1, ..., x_n)$

3.1 Metodi per trovare stimatori

3.1.1 Metodo dei momenti

Partiamo dal campione $X_1, ..., X_n$ iid $\mathcal{L}(X_i)$ funzione di $\vec{\theta}$

Momenti teorici: $\mathbb{E}[X] = \mu_1 \quad \mathbb{E}[X^2] = \mu_2 \quad \mathbb{E}[X^k] = \mu_k$ (quando esistono finiti)

Momenti campionari: $\frac{1}{n}\sum X_i = m_1 \quad \frac{1}{n}\sum X_i^2 = m_2 \quad \frac{1}{n}\sum X_i^k = m_k$

Oss. Di solito μ_j sono funzione di $\vec{\theta}$

Lo stimatore sarà quella statistica che ottengo uguagliando $\mu_j = m_j$ anche se μ_j sono valori e m_j sono VA

Def. $\widehat{F}_n(t) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{[X_i \leq t]}$ detta **funzione di ripartizione empirica**, so che $\widehat{F}_n(t) \to F_X(t)$

Oss. I momenti di F sono μ_j, mentre i momenti di $\widehat{F}_n(t)$ sono gli m_j

Teorema: (Di Glivenko-Cantelli) Come $\widehat{F}_n(t)$ stima F, allora m_j stimano i μ_j

Risolvendo in θ_j il sistema: $\begin{cases} m_1 = \mu_1(\vec{\theta}) \\ \vdots \\ m_h = \mu_h(\vec{\theta}) \end{cases}$ trovo gli stimatori $\widehat{\theta}_{j,MOM}$

Esempio: $X_1, ..., X_n \sim \mathcal{E}(\lambda)$

$$\mu_1 = \mathbb{E}[X] = \frac{1}{\lambda} = m_1 = \overline{X}_n \implies \widehat{\lambda}_{MOM} = \frac{1}{\overline{X}_n} = \frac{n}{\sum X_i}$$

Esempio: $X_1, ..., X_n \sim \mathcal{N}(\mu, \sigma^2)$ qui ho due incognite quindi dovrò usare almeno i primi due momenti

$$\begin{cases} \mu_1 = \mu = \overline{X}_n \\ \mu_2 = \mathbb{E}\left[X^2\right] = \mathrm{Var}[X] + \mathbb{E}[X]^2 = \theta_2 + \theta_1^2 = m_2 = \frac{1}{n}\sum X_i^2 \end{cases}$$

$$\text{Ottengo: } \begin{cases} \widehat{\mu}_{MOM} = \overline{X}_n \\ \widehat{\sigma}^2_{MOM} = \frac{1}{n}\sum X_i^2 - \overline{X}_n^2 \end{cases} \implies \begin{cases} \widehat{\mu} = \overline{X}_n \\ \widehat{\sigma}^2_{MOM} = \frac{1}{n}\sum (X_i - \overline{X}_n)^2 \end{cases}$$

Esempio: $X_1, ..., X_n \sim \mathcal{U}_{[-\theta,\theta]}$

$$\begin{cases} \mu_1 = 0 \\ \mu_2 = \frac{4}{12}\theta^2 = \frac{\theta^2}{3} \end{cases} \implies \frac{\theta^2}{3} = \frac{1}{n}\sum X_i^2 \implies \widehat{\theta}_{MOM} = \sqrt{\frac{3}{n}\sum X_i^2}$$

Valutiamo il metodo dei momenti

Pro: Sono equazioni algebriche, semplici da risolvere

Contro: Può succedere che le stime fornite da $\widehat{\theta}_{MOM}$ non siano ammissibili, ovvero $\widehat{\theta}_{MOM} \notin \Theta$ (esempio varianze negative)

Esempio: $X_1, ..., X_n \sim Bi(k, p) \quad \Theta = \mathbb{N} \times [0, 1]$

$$\begin{cases} \mu_1 = kp = \overline{X}_n \\ \mu_2 = kp(1-p) + (kp)^2 = \frac{1}{n}\sum X_i^2 \end{cases} \implies \begin{cases} \widehat{p} = \frac{\overline{X}_n}{\widehat{k}} \\ \overline{X}_n - \widehat{k}\frac{\overline{X}_n^2}{\widehat{k}^2} + \widehat{k}^2\left(\frac{\overline{X}_n}{\widehat{k}}\right)^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 \end{cases}$$

$$\implies \begin{cases} \widehat{p} = \frac{\overline{X}_n}{\widehat{k}} \\ \overline{X}_n + \overline{X}_n^2 - \frac{1}{n}\sum X_i^2 = \frac{\overline{X}_n^2}{\widehat{k}} \end{cases} \implies \begin{cases} \widehat{p} = \frac{\overline{X}_n}{\widehat{k}} \\ \widehat{k} = \frac{\overline{X}_n^2}{\overline{X}_n - \frac{1}{n}\sum (X_i - \overline{X}_n)^2} \end{cases}$$

Oss. É evidente che $\widehat{k}$ non appartenga ai naturali

Lezione 6 (01/03/2023)

3.1.2 Metodo basato sulla verosimiglianza

Definizione.

$L(\vec{\theta}, \vec{x}) = f(\vec{x}, \vec{\theta})$ **Likelihood** è la densità di $\vec{X}$ vista come funzione di $\vec{\theta}$

Oss. Cercheremo il θ che massimizza la L, ovvero lo scenario più probabile

Esempio: Estrazione con reimmissione di 3 palline da un'urna che contiene una proporzione p incognita di palline bianche e (1-p) di palline nere

Sapendo di aver estratto due palline bianche su 3, scegliere tra $p = \frac{1}{2}$ e $p = \frac{1}{3}$

Dovrò calcolare la probabilità di aver estratto 2b su 3, sapendo il valore di p e valutare la probabilità più alta

$$X_1, ..., X_n \sim Be(p) \quad \sum X_i = 2 \quad \mathcal{L}(\sum_{i=1}^{3} X_i) \sim Bi(3, p)$$

$$\mathbb{P}_{p=\frac{1}{2}}(\sum X_i = 2) = \binom{3}{2}\left(\frac{1}{2}\right)^2 \frac{1}{2} = \frac{3}{8} \quad \mathbb{P}_{p=\frac{1}{3}}(\sum X_i = 2) = \binom{3}{2}\left(\frac{1}{3}\right)^2 \frac{2}{3} = \frac{2}{9}$$

Osservo che $\frac{3}{8} > \frac{2}{9} \implies$ scelgo $p = \frac{1}{2}$

Oss. Ovviamente nel pratico non potremo tutte le volte calcolare la probabilità per ogni valore del parametro, ci servirà un metodo migliore

Definizione.

Stimatore di massima verosimiglianza MLE (maximum likelihood estimator) $\widehat{\theta}_{MLE}$ è il valore del parametro per cui $L(\theta, \vec{x})$ raggiunge il massimo $\quad \widehat{\theta}_{MLE}(\vec{x}) = \underset{\theta \in \Theta}{ArgSup}\, L(\theta, \vec{x})$

Valutiamo il metodo:

Pro: Per definizione le stime fornite da $\widehat{\theta}_{MLE}$ sono sempre ammissibili, dato che prendo solo i valori $\theta \in \Theta$

Contro: Niente garantisce che ci sia un Max assoluto o che sia unico

Niente garantisce che il massimo si possa scrivere in forma chiusa, ovvero scriverlo in forma esplicita e quindi dovrò ricorrere a metodi di massimizzazione numerica

Quando possibile, il metodo per trovare il massimo sarà:

$\begin{cases} \frac{\partial L}{\partial \theta_i} = 0 \quad \forall i = 1...k \end{cases}$ Risolvendo questo sistema si ottiene l'equazione di verosimiglianza

Spesso, per esempio quando ho un campione, poiché le variabili sono indipendenti, avrò:

$$L(\vec{\theta}, \vec{x}) = f(\vec{x}, \vec{\theta}) = \prod_{i=1}^{n} f(x_i, \vec{\theta})$$

Però dato che il logaritmo è monotona crescente, vale $\underset{\theta\in\Theta}{ArgSup}(log\, L) = \underset{\theta\in\Theta}{ArgSup}(L)$

Definisco il logLiklihood $l(\vec{\theta}, \vec{x}) = log\, L(\vec{\theta}, \vec{x})$ e risolvo $\begin{cases} \dfrac{\partial l}{\partial \theta_i} = 0 \quad \forall i = 1...k \end{cases}$

Inoltre è vantaggioso anche perché la derivata di un prodotto è brutta e la derivata della somma è bella

Esempio: $X_1, ..., X_n \sim Be(p)$

$$L(p, \vec{x}) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x} \mathbb{1}_{\{0,1\}}(x_i) = p^{\sum x_i}(1-p)^{n-\sum x_i} \prod_{i=1}^{n} \mathbb{1}_{\{0,1\}}(x_i)$$

$l(p, \vec{x}) = \sum x_i\ log(p) + (n - \sum x_i) log(1-p) + c(\vec{x})$ posso non considerare tutto ciò che non dipende da p

$$\frac{\partial l}{\partial p} = \frac{\sum X_i}{p} - \frac{(n - \sum X_i)}{1-p} = 0$$

(1) $0 < \sum X_i < n \quad \dfrac{\sum X_i}{p} = \dfrac{(n - \sum X_i)}{1-p} \Leftrightarrow \sum X_i - p\sum X_i = np - p\sum X_i \Leftrightarrow p = \dfrac{\sum X_i}{n} \implies \widehat{p}_{MLE} = \overline{X}_n$

(2) $\sum X_i = 0 \quad \dfrac{\partial l}{\partial p} = -\dfrac{n}{1-p} < 0 \implies \widehat{p}_{MLE} = 0 = \overline{X}_n$

(3) $\sum X_i = n \quad \dfrac{\partial l}{\partial p} = \dfrac{n}{p} > 0 \implies \widehat{p}_{MLE} = 1 = \overline{X}_n$

Quindi: $\widehat{p}_{MLE} = \overline{X}_n$

Esempio: $X_1, ..., X_n \sim \mathcal{U}_{[0,\theta]}$

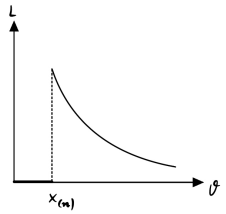
$$L(\theta, \vec{x}) = \frac{1}{\theta^n} \prod_{i=1}^{n} \mathbb{1}_{[0,\theta]}(x_i) = \frac{1}{\theta^n} \mathbb{1}_{[0,\theta]}(X_{(n)})$$

$$\mathbb{1}_{[0,\theta]}(X_{(n)}) \Leftrightarrow 0 \le X_{(n)} \le \theta \Leftrightarrow \mathbb{1}_{[X_{(n)},+\infty]}(\theta) \implies L(\theta, \vec{x}) = \frac{1}{\theta^n} \mathbb{1}_{[X_{(n)},+\infty]}(\theta)$$

In questo caso non serve fare calcoli, basta fare il disegno

Notiamo subito che il massimo della funzione L è in $X_{(n)}$

$\implies \widehat{\theta}_{MLE} = X_{(n)}$



Confrontiamo questo caso tra i due metodi visti: $\mu_1 = \dfrac{\theta}{2} \implies \widehat{\theta}_{MOM} = 2\overline{X}_n$, mentre $\widehat{\theta_{MLE}} = X_{(n)}$
In questo caso useremo $X_{(n)}$ perché è una statistica sufficiente, minimale e completa, quindi più comoda

Esempio: $X_1, ..., X_n \sim \mathcal{N}(\mu, \sigma^2)$

$k = 2 \;\; L(\mu, \sigma^2, \vec{x}) = \dfrac{1}{(2\sqrt{2\pi\sigma^2})^2}\, exp\left\{-\dfrac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right\} \implies l(\mu, \sigma^2, \vec{x}) = -\dfrac{n}{2}log(2\pi\sigma^2) - \dfrac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)$

Cerco i punti in cui la derivata di L cambia di segno e poi verifico se sono effettivamente massimi di L

$$\begin{cases} \dfrac{\partial l}{\partial \mu} = \dfrac{1}{\sigma^2}\sum(X_i-\mu) \\ \dfrac{\partial l}{\partial \sigma^2} = -\dfrac{n}{2}\dfrac{2\pi}{2\pi\sigma^2} + \dfrac{\sum(X_i-\mu)^2}{2(\sigma^2)^2} \end{cases} \qquad \begin{cases} \dfrac{\partial l}{\partial \mu} \geq 0 \Leftrightarrow \sum X_i \geq n\mu \Leftrightarrow \mu \leq \dfrac{\sum X_i}{n} \\ \dfrac{\partial l}{\partial \sigma^2} \geq 0 \Leftrightarrow \dfrac{\sum(X_i-\mu)^2}{\sigma^2} \geq n \Leftrightarrow \sigma^2 \leq \dfrac{\sum(X_i-\mu)^2}{n} \end{cases}$$

Nella 2° equazione ho il parametro μ, lo devo sostituire con il suo stimatore: $\begin{cases} \widehat{\mu}_{MLE} = \overline{X}_n \\ \widehat{\sigma^2}_{MLE} = \dfrac{\sum(X_i-\overline{X}_n)^2}{n} \end{cases}$

Proprietà di **invarianza** degli MLE dice che se $\widehat{\theta}_{MLE}$ è lo stimatore MLE di $\theta \implies \forall\tau$ funzione di θ,
lo stimatore MLE di $\tau(\theta)$ è $\tau(\widehat{\theta}_{MLE})$, ovvero $\widehat{\tau(\theta)}_{MLE} = \tau(\widehat{\theta}_{MLE})$

Esempio: $X_1, ..., X_n \sim \mathcal{P}(\lambda)$ trovare l'MLE di $\mathbb{P}(X_i = 0) = e^{-\lambda}$
Userò la proprietà di invarianza, cerco $\widehat{\lambda}_{MLE}$

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

$$\frac{\partial l}{\partial \lambda} = \sum_{i=1}^{n}\frac{\partial}{\partial\lambda}\left(log\left(\frac{e^{-\lambda}\lambda^{X_i}}{X!}\right)\right) = \sum_{i=1}^{n}\frac{\partial}{\partial\lambda}(-\lambda + X_i log\lambda) = \sum(-1 + \frac{X_i}{\lambda} = -n + \frac{\sum X_i}{\lambda})$$

$$\frac{\partial l}{\partial \lambda} \geq 0 \Leftrightarrow \lambda \leq \frac{\sum X_i}{n} \implies \widehat{\lambda}_{MLE} = \overline{X}_n$$

$$\implies \widehat{(e^{-\lambda})} = e^{-\overline{X}_n} \text{per principio invarianza}$$

Esempio: Abbiamo appena calcolato lo stimatore $\widehat{\sigma^2}_{MLE}$ per un campione gaussiano, allora vale:

$$\widehat{\sigma}_{MLE} = \sqrt{\widehat{\sigma^2}_{MLE}} = \sqrt{\frac{\sum_{i=1}^{n}(X_i-\overline{X}_n)^2}{n}}$$

Esempio: Sapppiamo che $\widehat{p}_{MLE} = \overline{X}_n$, allora

$$\sigma^2 = \tau(p) = p(1-p) \implies \widehat{\sigma^2}_{MLE} = \overline{X}_n(1-\overline{X}_n)$$

Restrizione del Range:

Se restringo lo spazio dei parametri da $\theta \in \Theta$ a $\theta \in \Theta_0 \subset \Theta$, allora $\widehat{\theta^0_{MLE}} = \underset{\theta\in\Theta_0}{ArgSup} L(\theta, \vec{x})$

Lezione 7 (03/03/2022)

3.2 Analisi degli stimatori

Definizione.

Il MSE (mean squared error) è l'errore quadratico medio di uno stimatore T per un parametro incognito θ ed è definito nel seguente modo: $MSE_\theta(T) = \mathbb{E}_\theta[(T-\theta)^2]$

$$MSE_\theta(T) = \mathbb{E}_\theta[(T - \mathbb{E}_\theta[T] + \mathbb{E}_\theta[T] - \theta)^2] = \mathbb{E}_\theta[(T-\mathbb{E}_\theta[T])^2 + (\mathbb{E}_\theta[T]-\theta)^2 + 2(T-\mathbb{E}_\theta[T])(\mathbb{E}_\theta[T]-\theta)] =$$

$$= \mathbb{E}_\theta[(T-\mathbb{E}_\theta[T])^2] + (\mathbb{E}_\theta[T]-\theta)^2 + 0 \implies MSE_\theta(T) = Var_\theta[T] + (\mathbb{E}_\theta[T]-\theta)^2$$

Definiamo la distorsione = bias := $\mathbb{E}_\theta[T] - \theta$

Se $\mathbb{E}_\theta[T] - \theta = 0$, allora T è detto non distorto per θ o unbiased

Oss. Tendenzialmente esiste un trade-off tra varianza e distorsione, per cui è difficile tenerli bassi entrambi

Dato $X_1, ..., X_n$ la media campionaria $\overline{X}_n$ è sempre uno stimatore non distorto della $\mathbb{E}[X_i]$
Infatti $\mathbb{E}\left[\overline{X}_n\right] = \mathbb{E}\left[\frac{1}{n}\sum X_i\right] = \frac{1}{n} n \mathbb{E}[X] = \mathbb{E}[X]$

Dato $X_1, ..., X_n \sim \mathcal{N}(\mu, \sigma^2)$ confrontiamo due stimatori della varianza
$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2 \qquad \widehat{\sigma}^2_{MLE} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2 = S_0^2$
Oss. Possiamo sfruttare il fatto che $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1) \sim \Gamma\left(\frac{n-1}{2}, \frac{1}{2}\right)$

$$MSE(S^2) = \text{Var}\left[S^2\right] + (\mathbb{E}\left[S^2\right] - \sigma^2)^2 = \left\{\left\{\mathbb{E}\left[S^2\right] = \frac{\sigma^2}{n-1}\mathbb{E}\left[\chi^2(n-1)\right] = \sigma^2\right\}\right\} = \text{Var}\left[S^2\right] =$$

$$= \frac{\sigma^4}{(n-1)^2}\text{Var}\left[\chi^2(n-1)\right] = \frac{\sigma^4}{(n-1)^2}(n-1)2 = \frac{2\sigma^4}{n-1}$$

$$S_0^2 = \frac{n-1}{n}S^2 \implies \text{Bias} = \mathbb{E}\left[S_0^2\right] - \sigma^2 = \frac{n-1}{n}\sigma^2 - \sigma^2 = -\frac{\sigma^2}{n} \quad \left(S_0^2 \text{ sottostima } \sigma^2\right)$$

$$\implies \text{Var}\left[S_0^2\right] = \frac{(n-1)^2}{n^2}\text{Var}\left[S^2\right] = \frac{2\sigma^4(n-1)}{n^2}$$

$$MSE(S_0^2) = \frac{2\sigma^4(n-1)}{n^2} + \frac{\sigma^4}{n^2} = \frac{(2n+1)\sigma^4}{n^2}$$

Confrontiamo i due MSE: $\frac{2\sigma^4}{n-1}$ $\frac{(2n+1)\sigma^4}{n^2}$

Equivale al confronto tra: $\forall\sigma^2$ $(2n-1)(n-1) = 2n^2 - 3n + 1$ e $2n^2$

E quindi: $1 - 3n$ e 0 Dove il primo è sempre miore del secondo

Qunidi $\forall\sigma^2 \quad MSE_{\sigma^2}[S_0^2] < MSE_{\sigma^2}[S^2]$ ovvero mi conviene prendere uno stimatore distorto

Questo perchè la varianza è sempre inversa all'errore, quindi conviene sovrastimare l'errore per essere cauti

Definiamo le seguenti relazioni tra stimatori e parametri:

Se $\mathbb{E}[T] < \theta$ sottostima

Se $\mathbb{E}[T] > \theta$ sovrastima

In generale tra due stimatori T_1 e T_2 di θ, scelgo lo stimatore con MSE inferiore

T_1 è preferibile a T_2 se $MSE_\theta(T_1) < MSE_\theta(T_2) \ \ \forall\theta$

Attenzione che il confronto tra MSE è un confronto tra due funzioni di θ, potrebbe succedere che la disuguaglianza tra MSE non sia vera $\forall\theta$

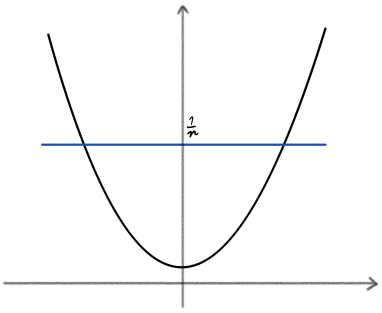
Ovvero che l'ordinamento del confronto tra MSE non è totale, potrebbero esserci stimatori non confrontabili

Esempio: $X_1, ..., X_n$ con $\mathcal{L}(x, \theta)$ che hanno $\theta = \mathbb{E}[X_i]$ e $1 = \text{Var}[X_i]$

$T_1 = \overline{X}_n \ \ T_2 = a\overline{X_n}$

$$MSE_\theta[T_1] = Var_\theta[T_1] = \text{Var}\left[\frac{1}{n}\sum X_i\right] = \frac{1}{n^2}\cdot n \cdot 1 = \frac{1}{n}$$

$$MSE_\theta[T_2] = \frac{a^2}{n} + (a\theta - \theta)^2 = \frac{a^2}{n} + (a-1)^2\theta^2$$



I due stimatori sono non confrontabili, però sceglierò la media campionaria, perché l'errore rimane "sotto controllo", mentre l'altro errore è quadratico

Esempio: $X_1 \sim \mathcal{P}(\lambda) \quad T(X_1) = (-1)^{X_1}$

$$\mathbb{E}[T] = \mathbb{E}\left[(-1)^{X_1}\right] = \sum_{k=0}^{+\infty}\left[(-1)^k \frac{e^{-\lambda}\lambda^k}{k!}\right] = e^{\lambda}\left[\sum_{k=0}^{+\infty}\frac{(-\lambda)^k}{k!}\right] = e^{-2\lambda}$$

T è uno stimatore non distorto di $e^{-2\lambda}$ $\qquad T = \begin{cases} 1 & \text{se } X_1 \text{ è pari} \\ -1 & \text{se } X_1 \text{ è dispari} \end{cases}$

Oss. Sto però stimando con -1 la funzione $e^{-2\lambda}$ che è sempre positivo e non potrà mai essere -1, inoltre posso ricavare che il MSE non è tanto basso

Oss. Questo controesempio mi dice che non è detto che uno stimatore unbiased sia buono. Proviamo a restringere il campo degli stimatori che cerchiamo

Definizione.

Per UMVUE si intente uniform minimum variance unbiased estimator, ovvero stimatore non distorto a varianza uniformemente minima (rispetto a θ)

T^* è **UMVUE** per θ se $\begin{cases} \mathbb{E}_\theta[T^*] = \theta \\ \mathbb{E}[T] = \theta \implies \text{Var}[T^*] \leq \text{Var}[T] \quad \forall\theta \quad \forall T \text{ stimatore non distorto di } \theta \end{cases}$

Teorema: Disuguaglianza di Cramer-Rao.

Siano $X_1, ..., X_n$ *iid* con $X_i \sim f(x_i, \theta)$ e $T(X_1, ..., X_n)$ stimatore di θ. Assumiamo che:

i) supporto di X_i non dipende da θ $\qquad$ ii) $\mathbb{E}\left[T^2\right] < +\infty$

iii) $\dfrac{d\mathbb{E}_\theta[T]}{d\theta} = \dfrac{d}{d\theta}\displaystyle\int_{\mathbb{R}^n} T(\vec{x}) f(\vec{x},\theta)\, d\vec{x} = \int_{\mathbb{R}} T(\vec{x}) \dfrac{\partial}{\partial\theta} f(\vec{x},\theta)\, d\vec{x}$

Allora $\quad 0 < \mathbb{E}_\theta\left[\left(\dfrac{\partial}{\partial\theta} log\, f(\vec{X},\theta)\right)^2\right] < +\infty \implies \mathrm{Var}\,[T] \geq \dfrac{\left[\dfrac{d}{d\theta}\mathbb{E}_\theta[T]\right]^2}{\mathbb{E}_\theta\left[\left(\dfrac{\partial}{\partial\theta} log\, f(\vec{X},\theta)\right)^2\right]}$

Oss. Nella famiglia esponenziale le condizioni sono rispettate

Oss. Inoltre se siamo nei non distorti, allora $\dfrac{d}{d\theta}\mathbb{E}_\theta[T] = 1$

Informazione di Fisher $= I_n(\theta) = \mathbb{E}_\theta\left[\left(\dfrac{\partial}{\partial\theta} log\, f(\vec{X},\theta)\right)^2\right]$

Oss. Detta informazione perché se ho I bassa significa che il limite sopra cui deve stare la varianza degli stimatori è alta e viceversa e quindi mi dice in anticipo che tipo di stimatori potrò trovare

Nella classe degli stimatori non distorti il limite di C.-R. è $\dfrac{1}{I_n(\theta)} \implies \mathrm{Var}\,[T] \geq \dfrac{1}{I_n(\theta)}$

Se io trovo uno stimatore T non distorto la cui varianza raggiunge il limite di C.-R. allora T è UMVUE

Però l'UMVUE può non raggiungere necessariamente il limite di C.-R.

Lezione 8 (07/03/2023)

Esempio: Trovaiamo l'informazione di Fisher per delle Poisson $\quad X_1, ..., X_n \sim \mathcal{P}(\lambda)$

$$I_n(\lambda) = \mathbb{E}\left[\left(\frac{\partial}{\partial\lambda} log\, f(\vec{X},\lambda)\right)^2\right]$$

$$\frac{\partial}{\partial\lambda} log\, f(\vec{X},\lambda) = \frac{\partial}{\partial\lambda} log\left[\prod_{i=1}^n \frac{e^{-\lambda}\lambda^{X_i}}{X!}\right] = \sum_{i=1}^n \frac{\partial}{\partial\lambda} log\left(e^{-\lambda}\lambda^{X_i}\right) = \sum_{i=1}^n \frac{\partial}{\partial\lambda}(-\lambda + X_i\, log\,\lambda) = \sum_{i=1}^n \left(-1 + \frac{X_i}{\lambda}\right) = \frac{n}{\lambda}\left(\overline{X}_n - \lambda\right)$$

$$\implies I_n(\lambda) = \mathbb{E}\left[\left(\frac{n}{\lambda}(\overline{X}_n - \lambda)\right)^2\right] = \frac{n^2}{\lambda^2}\mathbb{E}\left[\left(\overline{X}_n - \lambda\right)^2\right] = \frac{n^2}{\lambda^2}\mathrm{Var}\left[\overline{X}_n\right] = \frac{n^2}{\lambda^2}\cdot\frac{\lambda}{n} = \frac{n}{\lambda}$$

Oss. Spesso aiuta, nel calcolo di I, riportarsi a una varianza nota, come abbiamo fatto qui

Abbiamo quindi ottenuto che $\mathrm{Var}\,[T] \geq \dfrac{\left[\dfrac{d}{dx}\mathbb{E}\,[T]\right]^2}{I_n(\lambda)} = \dfrac{\lambda}{n}$ dato che lo stimatore è non distorto

Osservo che $\overline{X}_n$ è non distorto e $\mathrm{Var}\left[\overline{X}_n\right] = \dfrac{\lambda}{n}$ cioè raggiunge il limite di C.-R. $\implies \overline{X}_n$ è UMVUE per λ

Dim. (Cramer-Rao)

La dimostrazione di C.-R. è un'applicazione della disuguaglianza di Cauchy-Schwarz:

$$|\mathrm{Cov}\,(X,Y)\,|^2 \leq \mathrm{Var}\,[X]\,\mathrm{Var}\,[Y]$$

Per dimostrare C.-S. considero una va $aX+Y$

$$0 \leq \mathrm{Var}\,[aX+Y] = \underbrace{a^2\mathrm{Var}\,[X] + 2a\mathrm{Cov}\,(X,Y) + \mathrm{Var}\,[Y]}_{\text{Polinomio di ordine 2 in a}}$$

Ho quindi ottenuto una parabola che so essere sempre positiva e quindi ha determinate non positivo

$$\Delta = 4\mathrm{Cov}\,(X,Y)^2 - 4\mathrm{Var}\,[X]\,\mathrm{Var}\,[Y] \leq 0$$

$$\implies \mathrm{Var}\,[X] \geq \frac{|\mathrm{Cov}\,(X,Y)\,|^2}{\mathrm{Var}\,[Y]}$$

Si useranno come variabili $X = T$ e $Y = \dfrac{\partial}{\partial\theta}\, log\, f(\vec{X},\theta)$

Per l'ipotesi iii) del teorema vale: $\dfrac{d}{d\theta}\mathbb{E}_\theta[T] = \displaystyle\int_{\mathbb{R}^n} T(\vec{x})\frac{\partial}{\partial\theta} f(\vec{x},\theta)\, d\vec{x}$

$$\frac{\partial}{\partial\theta}\mathbb{E}_\theta[T] \overset{(3)}{=} \int_{\mathbb{R}^n} T(\vec{x})\cdot\frac{\partial}{\partial\theta} f(\vec{x},\theta)\, d\vec{x} = \int_{\mathbb{R}^n} T(\vec{x})\cdot\frac{\partial}{\partial\theta}\, log\, f(\vec{x},\theta)\cdot f(\vec{x},\theta)\, d\vec{x} = \mathbb{E}\,[T\cdot Y]$$

$$\text{Posto } T(\vec{X}) = 1 \text{ allora } 0 = \frac{d}{d\theta}\mathbb{E}\,[1] = \int_{\mathbb{R}^n}\frac{\partial}{\partial\theta}\, f(\vec{x},\theta)\, d\vec{x} = \int_{\mathbb{R}^n}\frac{\partial}{\partial\theta}\, log\, f(\vec{x},\theta)\cdot f(\vec{x},\theta)\, d\vec{x}$$

$$\implies \mathbb{E}\,[Y] = 0 \implies \mathrm{Cov}\,(X,Y) = \mathbb{E}\,[X\cdot Y] \text{ e } \mathrm{Var}\,[Y] = \mathbb{E}\left[Y^2\right]$$

Adesso per ricavare C.-R. dobbiamo applicare C.-S.

$$\mathrm{Var}\,[T] \geq \frac{\mathrm{Cov}\,(T,Y)^2}{\mathrm{Var}\,[Y]} = \frac{\left(\dfrac{d}{d\theta}\mathbb{E}_\theta[T]\right)^2}{\mathbb{E}\,[Y^2]} = \frac{\left[\dfrac{d}{d\theta}\mathbb{E}_\theta[T]\right]^2}{\mathbb{E}_\theta\left[\left(\dfrac{\partial}{\partial\theta}\, log\, f(\vec{x}\theta)\right)^2\right]}$$

□

Dalla dimostrazione si deduce che nella disuguaglianza C.-R. si raggiunge l'uguale quando
$X = T(\vec{x})$ è trasformazione lineare di $Y = \dfrac{\partial}{\partial\theta}\, log\, f(\vec{x},\theta)$

Ricerca dell'UMVUE

Teorema: Rao-Blackwell.

Siano T uno stimatore non distorto per θ e W una statistica sufficiente

Poniamo $M = \mathbb{E}_\theta[T|W]$, allora M è ancora uno stimatore non distorto per θ e $\text{Var}_\theta[M] \leq \text{Var}_\theta[T]$

Dim.

1) M è uno stimatore

$$M = \mathbb{E}_\theta[T(\vec{X})|W] = \int_{\mathbb{R}^n} T(\vec{x}) f(\vec{x}|W)\, d\vec{x}$$

$$\text{W è suff} \implies f(\vec{x}|W) \text{ non dipende da } \theta \implies \text{M stimatore}$$

2) M è non distorto

$$\mathbb{E}_\theta[M] = \mathbb{E}_\theta[\mathbb{E}_\theta[T|W]] = \mathbb{E}_\theta[T] = \theta$$

3) $\text{Var}[M] \leq \text{Var}[T]$

$$\text{Var}[T] = \text{Var}[\mathbb{E}_\theta[T|W]] + \underbrace{\mathbb{E}_\theta[\text{Var}[T|W]]}_{\geq 0 \text{ perché var} \geq 0} \implies \text{Var}[M] \leq \text{Var}[T]$$

□

Vogliamo far vedere che con W non sufficiente, non vale il teorema

Esempio: $X_1, X_2 \sim \mathcal{N}(\mu, 1) \quad T = \dfrac{X_1 + X_2}{2} \quad W = X_1$

$$\mathbb{E}[T|W] = \mathbb{E}\left[\overline{X}_2|X_1\right] = \frac{1}{2}\left[\mathbb{E}[X_1|X_1] + \mathbb{E}[X_2|X_1]\right] = \frac{1}{2}(X_1 + \mu) \quad \text{Non è uno stimatore}$$

Teorema: Unicità dell'UMVUE.

Sia $T(\vec{X})$ UMVUE per θ, allora T è unico

Dim.

Supponiamo che esista un altro T' UMVUE per θ

Ne costruisco un terzo. Sia $T^* = \dfrac{1}{2}(T + T')$

T^* è uno stimatore e ha $\mathbb{E}[T^*] = \dfrac{1}{2}(\mathbb{E}[T] + \mathbb{E}[T']) = \theta \implies$ anche T^* è non distorto

$$\text{Var}[T^*] = \frac{1}{4}\left[\text{Var}[T] + \text{Var}[T'] + 2\text{Cov}(T, T')\right] \overset{C.-S.}{\leq} \frac{1}{4}\left[\text{Var}[T] + \text{Var}[T'] + 2\sqrt{\text{Var}[T] \cdot \text{Var}[T']}\right]$$

Però dato che T e T' sono UMVUE allora devono avere stessa varianza $\implies \text{Var}[T^*] \leq \text{Var}[T]$ Ma dato che T è UMVUE e T^* è non distorto, in C.-S. vale l'uguaglianza $\implies \text{Var}[T^*] = \text{Var}[T]$

L'unico modo per cui questo è possibile è se $T' = aT + b$

$$\text{Var}[T] = \text{Cov}(T, T') = \text{Cov}(T, aT + b) = a\text{Var}[T] \implies a = 1$$

$$\mathbb{E}[T] = \mathbb{E}[T'] = \mathbb{E}[aT + b] = \mathbb{E}[T] + b \implies b = 0$$

$$\implies T = T' \implies \text{Se T esiste è unico}$$

□

Teorema: Lemma di Scheffè 2.

Sia T stimatore non distorto per θ (e quindi per ogni funzione $\tau(\theta)$)

Sia W stat suff, minimale e completa per θ

Allora $M = \mathbb{E}_\theta[T|W]$ è (l'unico) UMVUE per θ

Lezione 9 (13/03/20)

Dim.

M è uno stimatore, dato che W è sufficiente (visto in Cramer-Rau)

$$\mathbb{E}_\theta[M] = \mathbb{E}_\theta[\mathbb{E}_\theta[T|W]] = \mathbb{E}_\theta[T] = \theta$$

Supponiamo che M non sia UMVUE, ovvero che esista uno stimatore T' non distorto di θ tc $\text{Var}[T'] < \text{Var}[T]$

Usiamo teorema Rao-Blackwell su T'

$M' = \mathbb{E}[T'|W]$ è uno stimatore non distorto di θ con $\text{Var}[M'] \leq \text{Var}[T'] < \text{Var}[M]$

Osserviamo che M e M' sono funzioni di W e quindi $(M - M')$ è anch'essa funzione di W

$$\mathbb{E}[M - M'] = 0 \implies M - M' \text{ è una funzione } g(W) \text{ con media nulla, ma } W \text{ completo}$$

$$\implies \mathbb{P}(g(W) = 0) = 1 \implies M = M' \text{ qc}$$

Ma questo contraddice l'affermazione che M non fosse UMVUE, quindi lo è ed è unico

□

Esempio: $X_1, ..., X_n \sim Bi(k, \theta)$ con k fissato e θ incognito

Cerchiamo l'UMVUE per $\tau(\theta) = \mathbb{P}(X_i = 1) = k\theta(\theta)^{k-1}$

Sappiamo che $W = \sum X_j$ è stat suff min completa per un campione binomiale, dove $W \sim Bi(nk, \theta)$

Però $\mathbb{E}[W] = nk\theta \neq \tau(\theta)$, quindi mi devo trovare uno stimatore che condizionato a W mi dia τ

Quando $\tau(\theta)$ è una probabilità, posso prendere delle bernulli che abbiano p uguale a τ

$$Y_1, ..., Y_n \text{ tc } \; Y_i = \begin{cases} 1 & \text{se } X_i = 1 \\ 0 & \text{se } X_i \neq 1 \end{cases} \qquad \Longrightarrow \quad Y_i \sim \mathbb{1}_{X_i=1} \sim Be(\mathbb{P}(x_i = 1)) = Be\left(k\theta(1-\theta)^{k-1}\right)$$

Quindi $\overline{Y}_n = \dfrac{1}{n}\sum_{i=1}^{n} Y_i$ è stimatore non distorto di $\tau(\theta)$

Adesso devo fare l'attesa condizionata, che è la parte più "contosa":

$$\text{UMVUE per } \tau(\theta) \text{ è } \mathbb{E}\left[\overline{Y}_n \Big| \sum_{j=1}^{n} X_j\right] = \frac{1}{n}\sum_{j=1}^{n} \mathbb{E}\left[Y_i \Big| \sum X_j\right] \overset{\text{sono iid}}{=} \mathbb{E}\left[Y_1 \Big| \sum X_j\right]$$

$$\text{Però } \; \mathbb{E}\left[Y_1 \Big| \sum X_j = t\right] = \mathbb{P}\left(Y_1 = 1 \Big| \sum_{j=1}^{n} X_j = t\right) = \frac{\mathbb{P}(Y_1 = 1 \cap \sum X_j = t)}{\mathbb{P}(\sum X_j = t)} = \frac{\mathbb{P}(X_1 = 1 \cap \sum X_j = t)}{\mathbb{P}(\sum X_j = t)} =$$

Adesso gli eventi al numeratore non sono indipendenti, quindi non posso spezzare l'intersezione, però posso renderli indipendenti:

$$= \frac{\mathbb{P}\left(X_1 = 1 \cap \sum_{j=2}^{n} X_j = t-1\right)}{\mathbb{P}(\sum X_j = t)} = \frac{\mathbb{P}(X_1 = 1)\,\mathbb{P}\left(\sum_{j=2}^{n} X_j = t-1\right)}{\mathbb{P}(\sum X_j = t)} =$$

Dove $\quad X_j \sim Bi(k,\theta) \quad \sum_{j=2}^{n} X_j \sim Bi((n-1)k,\theta) \quad \sum_{j=1}^{n} X_j \sim Bi(nk,\theta)$

$$= \frac{k\theta(1-\theta)^{k-1}\binom{(n-1)k}{t-1}\theta^{t-1}(1-\theta)^{(n-1)k-t+1}}{\binom{nk}{t}\theta^{t}(1-\theta)^{nk-t}} = \frac{\binom{(n-1)k}{t-1}k}{\binom{nk}{t}} = \mathbb{E}\left[Y_1 \Big| \sum X_j = t\right]$$

Quindi l'UMVUE per $\; k\theta(1-\theta)^{k-1} \;$ è $\quad \dfrac{\binom{(n-1)k}{\sum X_j - 1}k}{\binom{nk}{\sum X_j}} = \mathbb{E}\left[\overline{Y}_n \Big| \sum X_j\right]$

3.3 Informazione di Fisher

Definizione.

Informazione di Fisher $= I_n(\theta) = \mathbb{E}_\theta\left[\left(\dfrac{\partial}{\partial\theta} log\, f(\vec{X},\theta)\right)^2\right]$

Lemma1:

Sia $X_1, ..., X_n$ iid, allora $I_n(\theta) = nI_1(\theta) = n\mathbb{E}\left[\left(\dfrac{\partial}{\partial\theta} log\, f(X_1,\theta)\right)^2\right]$

Dim. (Lemma1)

$$I_n(\theta) = \mathbb{E}_\theta \left[\left(\frac{\partial}{\partial \theta} log \prod_{i=1}^{n} f(X_i, \theta) \right)^2 \right] = \mathbb{E}_\theta \left[\left(\sum_{i=1}^{n} \frac{\partial}{\partial \theta} log\, f(X_i, \theta) \right)^2 \right]$$

Quando lavoriamo con l'informazione di Fisher le ipotesi per la disuguaglianza di Cramer-Rao le ho sempre, quindi posso scambiare integrale e derivata e avevamo visto che grazie a questo vale:

$$\mathbb{E}\left[\frac{\partial}{\partial \theta} log\, f(X_i, \theta) \right] = 0 \quad \text{quindi posso scrivere la media del quadrato con la varianza}$$

$$I_n(\theta) = \text{Var}\left[\sum_{i=1}^{n} \frac{\partial}{\partial \theta} log\, f(X_i \theta) \right] \overset{ind}{=} \sum_{i=1}^{n} \text{Var}\left[\frac{\partial}{\partial \theta} log\, f(X_i, \theta) \right] \overset{iid}{=} n\text{Var}\left[\frac{\partial}{\partial \theta} log\, f(X_i, \theta) \right] =$$

$$\text{Per quanto detto prima} \quad = n\mathbb{E}\left[\left(\frac{\partial}{\partial \theta} log\, f(X_i, \theta) \right)^2 \right] = nI_1(\theta)$$

□

Lemma2:

Se in aggiunta alle condizioni di Cramer-Rao si ha che: $\frac{\partial}{\partial \theta}\left[\int_{\mathbb{R}^n} \frac{\partial}{\partial \theta} f(\vec{x}, \theta)\, d\vec{x} \right] = \int_{\mathbb{R}^n} \frac{\partial^2}{\partial \theta^2} f(\vec{x}, \theta)\, dx$

Allora $I_n(\theta) = -\mathbb{E}_\theta \left[\frac{\partial^2}{\partial \theta^2} log\, f(\vec{x}, \theta) \right]$

Dim. (Lemma2)

$$\mathbb{E}\left[\frac{\partial^2}{\partial \theta^2} log\, f(\vec{x}, \theta) \right] = \int_{\mathbb{R}^n} \frac{\partial^2}{\partial \theta^2} (log\, f(\vec{x}, \theta)) \cdot f(\vec{x}, \theta)\, d\vec{x} = \int_{\mathbb{R}^n} \frac{\partial}{\partial \theta} \left(\frac{\frac{\partial}{\partial \theta} f(\vec{x}, \theta)}{f(\vec{x}, \theta)} \right) f(\vec{x}, \theta)\, d\vec{x} =$$

$$= \int_{\mathbb{R}^n} \frac{\partial^2}{\partial \theta^2} f(\vec{x}, \theta) \cdot \frac{1}{f(\vec{x}, \theta)} f(\vec{x}, \theta)\, d\vec{x} + \int_{\mathbb{R}^n} -\frac{1}{f^2(\vec{x}, \theta)} \cdot \left(\frac{\partial}{\partial \theta} f(\vec{x}, \theta) \right)^2 f(\vec{x}, \theta)\, dx$$

$$\text{Per l'ipotesi del lemma} \quad \int_{\mathbb{R}^n} \frac{\partial^2}{\partial \theta^2} f(\vec{x}, \theta)\, dx = \frac{\partial}{\partial \theta}\left[\int_{\mathbb{R}^n} \frac{\partial}{\partial \theta} f(\vec{x}, \theta)\, d\vec{x} \right] =$$

$$\frac{\partial}{\partial \theta}\left[\int_{\mathbb{R}^n} \frac{\partial}{\partial \theta} f(\vec{x}, \theta) \cdot \frac{1}{f(\vec{x}, \theta)} \cdot f(\vec{x}, \theta)\, d\vec{x} \right] = \frac{\partial}{\partial \theta}\left[\mathbb{E}\left[\frac{\partial}{\partial \theta} log\, f(\vec{x}, \theta) \right] \right] = 0$$

In conclusione, abbiamo ottenuto che:

$$\mathbb{E}\left[\frac{\partial^2}{\partial \theta^2} log\, f(\vec{x}, \theta) \right] = -\int_{\mathbb{R}^n} \frac{1}{f^2(\vec{x}, \theta)} \cdot \left(\frac{\partial}{\partial \theta} f(\vec{x}, \theta) \right)^2 f(\vec{x}, \theta)\, dx =$$

$$= -\int_{\mathbb{R}^n} \left(\frac{\partial}{\partial \theta} log\, f(\vec{x}, \theta) \right)^2 f(\vec{x}, \theta)\, d\vec{x} = -I_n(\theta)$$

□

Oss. Quindi se ho delle $log\, L$ (log verosimiglianze) "piatte", cioè hanno una concavità bassa, avrò il limite di CR molto alto (stimatori con varianza alta) che non ci piace

Invece in situazioni in cui la $log\, L$ è poco piatta, vuol dire che ha derivata seconda negativa grande (concavità grande) e quindi il limite di CR è basso

Esempio: Campione gaussiano ha la $logL = -\sum_{i=1}^{n}(x_i - \mu)^2$ che ha concavità alta e quindi ci piace

Teorema.

$X_1, ..., X_n$ iid con legge $f(X_i, \theta)$ appartenente EF, cioè $f(x,\theta) = h(x)x(\theta)\, exp\{w_1(\theta)t_1(x)\}$ tale che $\exists \frac{d}{d\theta} w_1(\theta) \neq 0$ e continua $\forall \theta$ e $\sum_{j=1}^{n} t_1(X_j)$ è sufficiente per θ

Allora le condizioni di CR sono soddisfatte e posto $T(\vec{X}) = \dfrac{\sum_{j=1}^{n} t_1(X_j)}{n}$, vale

$$\text{Var}\left[T(\vec{X})\right] = \frac{\left(\frac{d}{d\theta}\mathbb{E}_\theta\left[t_1(X)\right]\right)^2}{\mathbb{E}\left[\left(\frac{\partial}{\partial\theta} log\, f(\vec{X},\theta)\right)^2\right]}$$

Quindi $T(\vec{X})$ è UMVUE per $\mathbb{E}_\theta[t_1(X_j)]$

Lezione 10 (14/03/20)

Dim. (*)

Facciamo un'osservazione preliminare

$$0 = \frac{d}{d\theta}\int_{\mathbb{R}} f(x,\theta)\, dx = \frac{d}{d\theta}\int_{\mathbb{R}} h(x)c(\theta)\, exp\{w_1(\theta)t_1(x)\}\, dx =$$

Posso portare dentro all'integrale la derivata perchè valgono le ipotesi di CR

$$= \int_{\mathbb{R}} h(x)c'(\theta)exp\{w_1(\theta)t_1(x)\}\, dx + \int_{\mathbb{R}} h(x)c(\theta)exp\{w_1(\theta)t_1(x)\}\, w_1'(\theta)t_1(x)\, dx =$$

$$= \frac{c'(\theta)}{c(\theta)}\int_{\mathbb{R}} h(x)c(\theta)exp\{w_1(\theta)t_1(x)\}\, dx + w_1'(\theta)\int_{\mathbb{R}} h(x)c(\theta)exp\{w_1(\theta)t_1(x)\}\, t_1(x)\, dx =$$

Poichè $\int_{\mathbb{R}} h(x)c(\theta)exp\{w_1(\theta)t_1(x)\}\, dx = \int_{\mathbb{R}} f(x,\theta)\, dx = 1$ e $\dfrac{c'(\theta)}{c(\theta)} = \dfrac{d}{d\theta}\, log\, c(\theta)$

$$\implies 0 = \frac{d}{d\theta}\, log\, c(\theta) + w_1'(\theta)\mathbb{E}\left[t_1(x)\right] \implies \mathbb{E}_\theta[t_1(x)] = -\frac{1}{w_1'(\theta)} \cdot \frac{d}{d\theta}\, log\, c(\theta)$$

Ho così scritto la media in funzione di θ invece che in funzione di x

$$I_1(\theta) = \mathbb{E}_\theta\left[\left(\frac{\partial}{\partial\theta}\log f(x,\theta)\right)^2\right] = \mathbb{E}_\theta\left[\left(\frac{\partial}{\partial\theta}\left[\log h(x) + \log c(\theta) + w_1(\theta)t_1(x)\right]\right)^2\right] =$$

$$\mathbb{E}_\theta\left[\left(\frac{\partial}{\partial\theta}log\, c(\theta) + w_1'(\theta)t_1(x)\right)^2\right] = (w_1'(\theta))^2\,\mathbb{E}_\theta\left[\left(\frac{1}{w_1'(\theta)}\,\frac{\partial}{\partial\theta}log\, c(\theta) + t_1(x)\right)^2\right] =$$

$$= (w_1'(\theta))^2\,\mathbb{E}_\theta\left[(t_1(x) - \mathbb{E}\left[t_1(x)\right])^2\right] = (w_1'(\theta))^2\left[\mathbb{E}\left[(t_1(x))^2\right] - \mathbb{E}\left[t_1(x)\right]^2\right]$$

$$\implies I_1(\theta) = (w_1'(\theta))^2\,\mathrm{Var}\left[t_1(x)\right] \implies I_n(\theta) = n\,(w_1'(\theta))^2\,\mathrm{Var}\left[t_1(x)\right]$$

$$\frac{d}{d\theta}\mathbb{E}\left[t_1(x)\right] = \frac{d}{d\theta}\int_{\mathbb{R}} t_1(x)h(x)c(\theta)\,exp\{w_1(\theta)t_1(x)\}\,dx = ...$$

Con passaggi analoghi all'osservazione $= \dfrac{c'(\theta)}{c(\theta)}\displaystyle\int_{\mathbb{R}} t_1(x)f(x,\theta)\,dx + \int_{\mathbb{R}}(t_1(x))^2\,f(x,\theta)\cdot w_1'(\theta)\,dx =$

$$= \frac{d}{d\theta}\,log\, c(\theta)\mathbb{E}\left[t_1(x)\right] + w_1'(\theta)\mathbb{E}\left[(t_1(x))^2\right] = w_1'(\theta)\left[\mathbb{E}\left[\left(t_1(x)^2\right)\right] - \mathbb{E}\left[t_1(x)\right]^2\right] = w_1'(\theta)\mathrm{Var}\left[t_1(x)\right]$$

A questo punto abbiamo numeratore e denominatore del limite CR e mettendo insieme, otteniamo:

$$\text{Limite CR } = \frac{(w_1'(\theta))^2\,\mathrm{Var}\left[t_1(x)\right]^2}{n\,(w_1'(\theta))^2\,\mathrm{Var}\left[t_1(x)\right]} = \frac{\mathrm{Var}\left[t_1(x)\right]}{n} = \frac{1}{n^2}\cdot n\mathrm{Var}\left[t_1(x)\right] = \mathrm{Var}\left[\frac{1}{n}\sum_j t_1(X_j)\right]$$

Quindi abbiamo dimostrato che $T(X) = \dfrac{1}{n}\sum t_1(X_j)$ è UMVUE della sua media, ovvero $\mathbb{E}\left[t_1(X)\right]$

□

Esempio: $X_1, ..., X_n \sim \mathcal{P}(\lambda) \quad f(x,\lambda) = \dfrac{e^{-\lambda}\lambda^{-x}}{x!}\mathbb{1}_{\mathbb{N}}(x) = \dfrac{\mathbb{1}_{\mathbb{N}}(x)}{x!}\cdot e^{-\lambda}\,exp\{x\cdot log(\lambda)\}$

Qunidi $w_1'(x) = \dfrac{1}{\lambda}$ e sapendo che $\dfrac{1}{n}\displaystyle\sum_{j=1}^{n} t_1(X_j)$ è UMVUE per $-\dfrac{1}{w_1'(\lambda)}\cdot\dfrac{d}{d\lambda}\,log\, c(\lambda)$

Questo equivale a dire che per il campione gaussiano $\overline{X}_n$ è UMVUE per $-\dfrac{1}{\frac{1}{\lambda}}\cdot\dfrac{d}{d\lambda}\,log\left(e^{-\lambda}\right) = \lambda$

L'informazione di Fisher calcolata usando la legge del campione, coincide con quella calcolata usando la legge di una T statistica sufficiente

Questo perché avevamo visto che $f(\vec{x},\theta) = h(\vec{x})g(T(\vec{x}),\theta)$

$$\implies log\, f(\vec{x},\theta) = log\, h(\vec{x}) + log\, g(T(\vec{x}),\theta) \implies \frac{\partial}{\partial\theta}\,lof\, f(\vec{x},\theta) = \frac{\partial}{\partial\theta}\,log\, g(T(\vec{x}),\theta)$$

E quindi nell'informazione di Fisher le due cose sono uguali

3.4 Pillole sull'approccio bayesiano

Noi avevamo visto θ come un parametro incognito, per l'approccio Bayesiano θ è vista come una variabile aleatoria, perché il suo valore non è fissato a priori

La legge di θ si chiamna Prior e viene indicata con $\Pi(\theta)$
Quindi quando scrivo la legge di $\vec{X}$ la scriverò $f(\vec{X}|\theta)$ perché è condizionata a una VA

Quando faccio inferenza, calcolo la legge di $(\theta|\vec{x})$ e quindi la legge, detta a posterior $\Pi(\theta|\vec{x})$
Per il teorema di Bayes, ho: $\Pi(\theta|\vec{X}) = \dfrac{f(\vec{x}|\theta)\cdot\Pi(\theta)}{m(\vec{x})} = \dfrac{f(\vec{x}|\theta)\cdot\Pi(\theta)}{\displaystyle\int_\Theta f(\vec{x}|\theta)\Pi(\theta)\,d\theta}$

Esempio: Modello beta-binomiale

$$X_i|\theta \sim Be(\theta) \qquad \Pi(\theta) \sim Beta(\alpha,\beta)$$

$$\mathcal{L}(\textstyle\sum X_i|\theta) \sim Bi(n,\theta) \qquad \Pi(\theta|\sum X_i) \sim Beta(\alpha+\sum X_i\,,\, n-\sum X_i+\beta)$$

Per fare inferenza serve calcolare $\mathbb{E}\left[\Pi(\theta|\sum X_i)\right]$
Quindi scrivo la prior (a intuito) e poi fa l'esperimento e trova la condizionata $(\theta|\sum X_i)$ e trova la stima del parametro come la media della postirior:

$$\Pi(\theta) \sim Beta(\alpha,\beta) \qquad \widehat{\theta}_{\text{Bayes}} = \frac{\alpha+\sum X_i}{n+\alpha+\beta} = \frac{n}{\alpha+\beta+n}\left(\frac{\sum X_i}{n}\right) + \frac{\alpha+\beta}{\alpha+\beta+n}\left(\frac{\alpha}{\alpha+\beta}\right)$$

Ma riscritta così possiamo vedere che la stima è una media convessa (ponderata) della media che avevo a priori e della media campionaria e vediamo che al crescere di n l'informazione della prior perde di valore e torniamo alla media campionaria

Oss. L'idea di approcciare θ come VA è molto elegante, però i modelli per cui si riescono ad affrontare questi conti sono pochi, per gli altri bisogna procedere con simulazioni. Inoltre rimane il grosso problema della scelta della prior

4 Test d'ipotesi

Per test d'ipotesi si intende la verifica delle ipotesi statistiche

Definizione.

L'**ipotesi statistica** è un'affermazione su parametri incogniti $\vec{\theta}$ della legge di $\vec{X}$

Definizione.

In un problema di test d'ipotesi si introducono due ipotesi complementari: L'ipotesi nulla H_0 e l'ipotesi alternativa H_1

Oss. L'ipotesi statistica è la formulazione delle ipotesi nulla e alternativa, mentre il test d'ipotesi è un processo decisionale che ci porta a scegliere tra H_0 e H_1

Queste due ipotesi sono dette complementari perché se H_0 è tc $\vec{\theta} \in \Theta_0$,
Allora definiremo H_1 tc $\vec{\theta} \in \Theta_0^C$, ovvero sta nel complementare

Definizione.

Un **test d'ipotesi** è una regola che specifica:
a) Per quali valori di $\vec{x}$ accetto H_0
b) Per quali valori di $\vec{x}$ rifiuto H_0 (accetto H_1)

Dovrò definire la regione critica $\quad RC = \{\vec{x} \in \mathbb{R}^n$ tc decido di rifiutare $H_0\}$
RC viene specificata sulla base dei valori assunti da una statistica test $W(\vec{x})$

Oss. Nella definizione della regione critica, che è ciò che ci porta alla decisione, deve esserci solo il campione e la statistica, non può esserci il parametro incognito

Esempio: Abbiamo delle bottigliette con scritto $500ml$, però vogliamo vedere se la media è più bassa.
Faremo dei tentativi per valutare se lo è veramente: $\quad X_1, ..., X_n \sim \mathcal{N}(\mu, \sigma^2)$
$H_0 : \mu \geq 500ml \qquad H_1 : \mu < 500ml$
Quindi definisco la regione critica come $\quad RC : \{\overline{X}_n < 500 - k\}$

Oss. Attenzione che non avremmo potuto definire RC con μ al posto di $\overline{X}_n$
Oss. Tendenzialemnte H_0 è l'ipotesi "vera fino a prova contraria"

Lezione 11 (20/03/20)

Vediamo dei metodi per costruire dei test d'ipotesi e dei metodi per valutarli e confrontarli

4.1 Test del rapporto di verosimiglianza

Detto anche LRT, cioè Likelihood ratio test

Date delle ipotesi $H_0 : \theta \in \Theta_0 \qquad H_1 : \theta \in \Theta_0^C$, definiamo il LRT

Definizione.

La statistica test per il LRT è $\lambda(\vec{x}) = \dfrac{\sup\limits_{\theta\in\Theta_0} L(\theta, \vec{x})}{\sup\limits_{\theta\in\Theta} L(\theta, \vec{x})} = \dfrac{L(\widehat{\theta}^{\,0}_{MLE})}{L(\widehat{\theta}_{MLE})}$

La regione critica è $RC = \{\lambda(\vec{x}) \leq c\}$ con $0 \leq c \leq 1$

Oss. $0 \leq \lambda(\vec{x}) \leq 1$ perché sappiamo che sia numeratore che denominatore sono positivi e il sup in un insieme più grande è sicuramente maggiore uguale

Oss. Ha senso porre la regione critica con $\leq$ perché se λ è piccolo, vuol dire che il sup in Θ_0 è molto più piccolo del sup in Θ, ma quindi è molto probabile che il parametro appartenga a Θ_0^C

Esempio: $X_1, ..., X_n \sim \mathcal{N}(\theta, 1)$ $\begin{cases} H_0 : \ \theta = \theta_0 \\ H_1 : \ \theta \neq \theta_0 \end{cases}$

Oss. Al numeratore di λ scriviamo direttamente θ_0 perché è il sup in un punto

$$\lambda(\vec{x}) = \frac{L(\theta_0, \vec{x})}{\sup_{\theta\in\Theta} L(\theta, \vec{x})} = \frac{L(\theta_0, \vec{x})}{L(\widehat{\theta}_{MLE}, \vec{x})} = \frac{L(\theta_0, \vec{x})}{L(\overline{x}_n, \vec{x})}$$

$$\lambda(\vec{x}) = \frac{\left(\frac{1}{2\pi}\right)^{\frac{n}{2}} exp\left\{-\frac{1}{2}\sum_{i=1}^{n}(x_i - \theta_0)^2\right\}}{\left(\frac{1}{2\pi}\right)^{\frac{n}{2}} exp\left\{-\frac{1}{2}\sum_{i=1}^{n}(x_i - \overline{x}_n)^2\right\}} = exp\left\{\frac{1}{2}\left[-\sum_{i=1}^{n}(x_i - \theta_0)^2 + \sum_{i=1}^{n}(x_i - \overline{x}_n)^2\right]\right\}$$

Oss. $\sum\limits_{i=1}^{n}(x_i - \theta_0)^2 = \sum\limits_{i=1}^{n}(x_i - \overline{x}_n + \overline{x}_n - \theta_0)^2 = \sum\limits_{i=1}^{n}(x_i - \overline{x}_n)^2 + \sum\limits_{i=1}^{n}(\overline{x}_n - \theta_0)^2 + 0$

$$\implies \lambda(\vec{x}) = exp\left\{-\frac{n}{2}(\overline{x}_n - \theta_0)^2\right\}$$

$$RC = \left\{ exp\left\{ -\frac{n}{2}(\overline{X}_n - \theta_0)^2 \right\} \leq c \right\} = \left\{ \left(\overline{X}_n - \theta_0\right)^2 \geq -\frac{2\,log\,c}{n} \right\} = \left\{ \vec{x} \ : \ |\overline{X}_n - \theta_0| \geq \sqrt{-\frac{2\,log\,c}{n}} \right\}$$

Oss. Questa regione critica ha senso perché equivale a valutare la distanza tra θ_0 e $\overline{X}_n$ e se è maggiore di un certo valore allora rifiuto e questo ha senso perchè supponendo che θ_0 sia vera, allora la media campionaria è disposta come una gaussiana centrata in θ_0

Esempio: $X_1, ..., X_n \sim \theta + \mathcal{E}(1) \qquad f(x, \theta) = exp\{-(x-\theta)\}\mathbb{1}_{[0,+\infty]}(x)$

$$L(\theta, \vec{x}) = \prod_{i=1}^{n} exp\{-(x_i - \theta)\}\mathbb{1}_{[0,\infty]}(x_i) = exp\left\{ -\sum x_i + n\theta \right\} \mathbb{1}_{(-\infty, X_{(1)}]}(\theta)$$

Guardando la funzione, concludo che $\widehat{\theta}_{MLE} = X_{(1)}$ +++grafico

Test: $\begin{cases} H_0 : \ \theta \leq \theta_0 \\ H_1 : \ \theta > \theta_0 \end{cases}$

$$\lambda(x) = \frac{\sup_{\theta \leq \theta_0} L(\theta, \vec{x})}{\sup_{\theta \in \mathbb{R}} L(\theta, \vec{x})} = \frac{\sup_{\theta \leq \theta_0} L(\theta, \vec{x})}{L(X_{(1)}, \vec{x})}$$

+++grafici

Dato che la curva di L è crescente

Se $X_{(1)} \leq \theta_0 \implies \sup_{\theta \leq \theta_0} L(\theta, \vec{x}) = L(X_{(1)}, \vec{x})$

Se invece $X_{(1)} > \theta_0 \implies \sup_{\theta \leq \theta_0} L(\theta, \vec{x}) = L(\theta_0, \vec{x})$

$$\lambda(\vec{x}) = \begin{cases} 1 \quad X_{(1)} \leq \theta_0 \\ \dfrac{L(\theta_0, \vec{x})}{L(X_{(1)}, \vec{x})} \quad X_{(1)} > \theta_0 \end{cases} = \begin{cases} 1 \quad X_{(1)} \leq \theta_0 \\ exp\{-n(X_{(1)} - \theta_0)\} \quad X_{(1)} > \theta_0 \end{cases}$$

$$RC = \{e^{-n(X_{(1)} - \theta_0)} \leq c\} = \{X_{(1)} \geq \theta_0 - \frac{log\,c}{n}\}$$

Teorema.

Siano $T(\vec{X})$ stat suff per θ e $\lambda^*(t), \lambda(\vec{x})$ le statistiche LRT basate su T e su $\vec{X}$

$$\text{Allora} \quad \lambda^*(T(\vec{x})) = \lambda(\vec{x}) \ \ \forall \vec{x}$$

Dim.

$$\lambda(\vec{x}) = \frac{\sup_{\theta\in\Theta_0} L(\theta,\vec{x})}{\sup_{\theta\in\Theta} L(\theta,\vec{x})} \overset{\text{fattorizzazione}}{=} \frac{\sup_{\theta\in\Theta_0} g(T(\vec{x}),\theta)\cancel{h(\vec{x})}}{\sup_{\theta\in\Theta} g(T(\vec{x}),\theta)\cancel{h(\vec{x})}} =$$
$$= \frac{\sup_{\theta\in\Theta_0} L^*(T(\vec{x}),\theta)}{\sup_{\theta\in\Theta} L^*(T(\vec{x}),\theta)} = \lambda^*(T(\vec{x}))$$

□

Oss. Adesso che abbiamo capito la LRT facciamo considerazioni su un test in generale

4.2 Considerazioni di un test

Prendiamo un test con una RC e con le ipotesi $\begin{cases} H_0 : \ \theta \in \Theta_0 \\ H_1 : \ \theta \in \Theta_0^C \end{cases}$

	Accetto H_0	Rifiuto H_0
H_0 vero	OK	Errore di I tipo
H_1 vero	Errore di II tipo	OK

Commettere un errore di I tipo è peggio di commettere un errore di II tipo

$$\mathbb{P}_\theta(\vec{X} \in RC) = \begin{cases} \text{probabilità di commettere un errore di I tipo} \quad \text{se } \theta \in \Theta_0 \\ 1 - \mathbb{P}_\theta(\vec{X} \notin RC) = 1 - \text{probabilità di commettere un errore di II tipo} \quad \text{se } \theta \in \Theta_0^C \end{cases}$$

Definizione.

La **funzione potenza** di un test con regione critica RC è:

$$\beta(\theta) : \Theta \to [0,1] \qquad \beta(\theta) = \mathbb{P}_\theta(\vec{X} \in RC)$$

La funzione potenza "ideale", sarebbe $\beta(\theta) = \begin{cases} 0 & \text{se } \theta \in \Theta_0 \\ 1 & \text{se } \theta \in \Theta_0^C \end{cases}$

Oss. É impossibile riuscire ad avere una funzione così, perché l'unico modo per avere errore I =0 sarebbe quello di accettare sempre H_0, ma questo comporterebbe errore II = 1

Esempio: Dati la VA e i test seguenti $X \sim Bi(5,\theta)$ $\begin{cases} H_0: \ \theta \leq \frac{1}{2} \\ H_1: \ \theta > \frac{1}{2} \end{cases}$

Voglio confrontare le funzioni potenza di:

Test1: $RC_1 = \{X = 5\}$

Test2: $RC_2 = \{X \geq 3\}$

$\beta_1(\theta) = \mathbb{P}_\theta(X = 5) = \theta^5$

$\beta_2(\theta) = \mathbb{P}_\theta(X \geq 3) = \binom{5}{3}\theta^3(1-\theta)^2 + 5\,\theta^4(1-\theta) + \theta^5$

+++grafico

Le due funzioni sono crescenti, quindi per entrambe vale che $\mathbb{P}\,(\text{errore I tipo}) \leq \mathbb{P}_{\theta_0}(RC) = \beta(\theta_0)$

$\beta_1(\theta_0) = \left(\frac{1}{2}\right)^5 = 0,03125$ che è un buon limite

$\beta_1(\theta) \geq 0,8 \implies \theta \geq 0,96$ quindi questo test va bene in Θ_0, ma fuori non è molto forte

$\beta_2(\theta_0) = (10+5+1)\left(\frac{1}{2}\right)^5 = \frac{1}{2}$ che è troppo alta

Lezione 12 (21/03/2023)

Esempio: $X_1, ..., X_n \sim \mathcal{N}(\mu, \sigma^2)$ con σ^2 noto

LRT per $\begin{cases} H_0 \ \ \mu \leq \mu_0 \\ H_1 \ \ \mu > \mu_0 \end{cases}$ con $RC = \left\{ \frac{\overline{X}_n - \mu_0}{\frac{\sigma}{\sqrt{n}}} > c \right\}$

$$\beta(\mu) = \mathbb{P}_\mu\left(\frac{\overline{X}_n - \mu_0}{\frac{\sigma}{\sqrt{n}}} > c\right) = \mathbb{P}_\mu\left(\underbrace{\frac{\overline{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}}}_{Z} > c + \frac{\mu_0 - \mu}{\frac{\sigma}{\sqrt{n}}}\right) = 1 - \Phi\left(c + \frac{\mu_0 - \mu}{\frac{\sigma}{\sqrt{n}}}\right)$$

Al crescere di μ, $\frac{\mu_0 - \mu}{\frac{\sigma}{\sqrt{n}}}$ decrcesce, Φ decresce e quindi β è crescente in μ

$$\lim_{\mu \to -\infty} \beta(\mu) = 0 \qquad \lim_{\mu \to +\infty} \beta(\mu) = 1 \qquad \beta(\mu_0) = 1 - \Phi(c)$$

Inoltre per modificare la crescita di questa curva, l'unica cosa che possiamo fare è modificare n

Al crescere di n la funzione cresce più velocemente

Vogliamo controllare (limitare) la massima probabilità di commettere un errore di I tipo

$$\sup_{\mu \leq \mu_0} \beta(\mu) = \beta(\mu_0)$$

$$\beta(\mu_0) = 0.10 \iff 1 - \Phi(c) = 0.10 \iff c = Z_{0.9} = 1.28$$

Vogliamo che la funzione potenza sia ≥ 0.8 per $\mu \geq \mu_0 + \sigma$, ovvero che oltre a quel punto la probabilità di commettere un errore II tipo sia minore del 20%, per far ciò devo imporre:

$$0.8 = \beta(\mu_0 + \sigma) = 1 - \Phi\left(1.28 - \frac{\sigma}{\frac{\sigma}{\sqrt{n}}}\right) = 1 - \Phi(1.28 - \sqrt{n}) \iff \Phi(1.28 - \sqrt{n}) = 0.2$$

$$\iff 1.28 - \sqrt{n} = Z_{0.2} = -0.84 \iff n = 4.49 \iff n \geq 5$$

++++GRAFICO

Definizione.

$\forall\, 0 \leq \alpha \leq 1$ un test con funzione potenza $\beta(\theta)$ è detto di **dimensione** α se il $\sup_{\theta \in \Theta_0} \beta(\theta) = \alpha$

Definizione.

$\forall\, 0 \leq \alpha \leq 1$ un test con funzione potenza $\beta(\theta)$ è detto di **livello** α se il $\sup_{\theta \in \Theta_0} \beta(\theta) \leq \alpha$

Oss. Queste definizioni si usano per controllare gli errori di I tipo

Esempio: $X_1, ..., X_n$ iid $\quad f(X, \theta) = e^{-(x,\theta)} \mathbb{1}_{[\theta, +\infty)} \quad$ con $\quad \begin{cases} H_0 : \ \theta \leq \theta_0 \\ H_1 : \ \theta > \theta_0 \end{cases}$

Avevamo trovato LRT con $RC = \left\{ X_{(1)} \geq \theta_0 - \frac{log\, c}{n} \right\}$

LRT serve per trovare una "forma" di regione critica, quello che poi dobbiamo fare è imporre il livello

Voglio un test LRT di livello α, ovvero $\quad \sup_{\theta \leq \theta_0} \beta(\theta) = \sup_{\theta \leq \theta_0} \mathbb{P}_\theta(RC) \leq \alpha$

$$\beta(\theta) = \mathbb{P}_\theta \left(X_{(1)} \geq \theta_0 - \frac{log\, c}{n} \right)$$

Avevamo trovato che $X_i = \theta T_i \ \ \text{con} \ \ T_i \sim \mathcal{E}(1)$

$$\implies minX_i = min(\theta + T_i) = \theta + T_{(1)} \qquad T_{(1)} \sim \mathcal{E}(n)$$

$$\mathbb{P}(X_{(1)} \geq k) = e^{-n(k-\theta)} \qquad \text{questa funzione è crescente in } \theta$$

$$\beta(\theta) = exp\left\{-n\left(\theta_0 - \frac{log\, c}{n} - \theta\right)\right\} \qquad \alpha = \sup_{\theta \leq \theta_0} \beta(\theta) = \beta(\theta_0) = exp\left\{-n\left(\theta_0 - \frac{log\, c}{n} - \theta_0\right)\right\} = c$$

Definizione.

Un test con funzione potenza $\beta(\theta)$ è detto **non distorto** se $\beta(\theta') \leq \beta(\theta'') \quad \forall \theta' \in \Theta_0^C \quad \forall \theta'' \in \Theta_0$

Definizione.

Sia C una classe di test per verificare $H_0: \ \theta \in \Theta_0$ VS $H_1: \ \theta \in \Theta_0^C$

Allora un test in C con funzione potenza $\beta(\theta)$ è **UMP**, uniformly most powerful,

se $\beta(\theta) \geq \beta'(\theta) \ \ \forall \theta \in \Theta_0^C$ e $\forall \beta'(\theta)$ funzione potenza di test in C

Tipicamente C è la classe funzione dei test di livello α

Oss. L'idea è che sono sicuro di non avere problemi in Θ_0 perché ho chiesto che la classe sia di livello α, quindi la β più forte è quella più forte in Θ_0^C

Oss. Il seguente lemma parte da una situazione più facile e generalizza in seguito

Teorema: Lemma di Neymann-Pearson.

Usando le "ipotesi semplici", ovvero $H_0: \ \theta = \theta_0$ VS $H_1: \ \theta = \theta_1$

Per β ci serviranno solamente $f(\vec{x}, \theta_i)$ per $i = 0, 1$

Sia RC tale che:

1) $\vec{x} \in RC$ se $f(\vec{x}, \theta_1) > kf(\vec{x}, \theta_0)$ e $\vec{x} \in RC^C$ se $f(\vec{x}, \theta_1) < kf(\vec{x}, \theta_0)$ con $k \geq 0$

2) $\alpha = \mathbb{P}_{\theta_0}(\vec{x} \in RC)$

Allora:

a) Qualunque test che soddisfa 1) e 2) è UMP di livello e dimensione α

b) Se esiste un test che soddisfa 1) e 2) con $k > 0$

Allora ogni test UMP di livello α, ha anche dimensione α (soddisfa la 2))

E soddisfa la 1) tranne che su un insieme A tc $\mathbb{P}_{\theta_0}(\vec{x} \in A) = \mathbb{P}_{\theta_1}(\vec{x} \in A) = 0$

Oss. Quindi poste le ipotesi semplici, che mi fanno scegliere tra due valori del parametro

Questo lemma si basa su un'idea simile alle "classi di equivalenza dei test" e dice che un test con 1) e 2) è UMP e se invece prendo un test UMP, allora sarà "quasi" uguale al test con 1) e 2)

Dim. (Neymann-Pearson*)

La 2) equivale a dire che il test è di dimensione α

Prendiamo $\Phi(\vec{x})$ una funzione test $\quad \Phi(\vec{x}) = \mathbb{1}_{RC}(\vec{x})$

a)

Siano $\Phi(\vec{x})$ la funzione test di un test che soddisfa 1) e 2) con funzione potenza $\beta(\theta)$

E $\Phi'(\vec{x})$ la funzione test di un altro test di livello α con $\beta'(\theta)$

$$(\Phi(\vec{x}) - \Phi'(\vec{x})) \cdot (f(\vec{x}, \theta_1) - kf(\vec{x}, \theta_0)) \geq 0 \ \ \forall \vec{x}$$

Se $\vec{x} \in RC$, definita da 1), allora questa diventa:

$$(1 - \Phi'(\vec{x}))\underbrace{(f(\vec{x}, \theta_1) - kf(\vec{x}, \theta_0))}_{\geq 0}$$

Se $\vec{x} \in RC^C$, per la 1)

$$(0 - \Phi'(\vec{x}))\underbrace{f(\vec{x}, \theta_1 - f(\vec{x}, \theta_0))}_{\leq 0}$$

Avevamo ottenuto che:

$$\int_{\mathbb{R}^n} (\Phi(\vec{x}) - \Phi'(\vec{x}))(f(\vec{x}, \theta_1) - kf(\vec{x}, \theta_0))\, d\vec{x} \geq 0$$

Poichè $\displaystyle\int_{\mathbb{R}^n} \mathbb{1}_B(\vec{x})\, d\vec{x} = \mathbb{P}(\vec{x} \in B)$

$$0 \leq \mathbb{P}_{\theta_1}(RC) - \mathbb{P}_{\theta_1}(RC') - k\mathbb{P}_{\theta_0}(RC) + k\mathbb{P}_{\theta_0}(RC') = \beta(\theta_1) - \beta'(\theta_1) - k\big(\beta(\theta_0) - \beta'(\theta_0)\big)$$

So che Φ soddisfa la 2) $\implies \beta(\theta_0) = \alpha$

E so che Φ' è di livello α $\implies \beta'(\theta_0) \leq \alpha$

$\implies \beta(\theta_0) - \beta'(\theta_0) \geq 0 \implies \beta(\theta_1) - \beta'(\theta_1) \geq 0 \implies$ Φ è UMP, oltre che essere di livello e dimensione α

b)

Sia Φ' la funzione test di test UMP di livello α, ma da a) so che Φ (che soddisfa 1) e 2)) è UMP di liv α

$$\implies \beta(\theta_1) = \beta'(\theta_1)$$

So che $\beta'(\theta_0) \leq \alpha$ e da a) so che $0 \leq -k(\beta(\theta_0) - \beta'(\theta_0)) \implies 0 \geq (\alpha - \beta'(\theta_0)) \implies \beta'(\theta_0) \geq \alpha$

$\implies \beta'(\theta_0) = \alpha$ quindi soddisfa 2)

$$\beta(\theta_1) - \beta'(\theta_1) - k(\beta(\theta_0) - \beta'(\theta_0)) = \int_{\mathbb{R}^n} (\Phi(\vec{x}) - \Phi(\vec{x}))(f(\vec{x}, \theta_1) - kf(\vec{x}, \theta_0))\, d\vec{x}$$

$\implies \Phi = \Phi'$ tranne che su un insieme A con $\mathbb{P}_{\theta_0}(\vec{x} \in A) = \mathbb{P}_{\theta_1}(\vec{x} \in A)$ □

Lezione 13 (27/03/23)

Oss. $\begin{cases} H_0 : f_0(\vec{x}) \\ H_1 : f_1(\vec{x}) \end{cases}$ non devono essere della stessa famiglia parametrica

Esempio: 13.1

Corollario:

Dato $\begin{cases} H_0 : \theta = \theta_0 \\ H_1 : \theta = \theta_1 \end{cases}$ sia $T(\vec{X})$ una statistica sufficiente per θ e $g(t, \theta)$ la legge di $T(\vec{X})$

Allora il test UMP di livello α basato su T è quello con $RC = \{g(t, \theta_1) > kg(t, \theta_2)\}$

Oss. Vale perché $f(\vec{x}, \theta) = g(t, \theta)h(\vec{x})$

Esempi: 13.2 13.3 13.4

Definizione.

Data una famiglia di leggi $\{f(x, \theta),\ \theta \in \Theta\}$, la famiglia è detta a MLR, o likelihood ratio monotona (non decrescente) se $\dfrac{f(x, \theta_1)}{f(x, \theta_2)}$ è monotona non decrescente in x $\forall \theta_1 > \theta_2$

Esempi: 13.5 13.6

Teorema: Karlin-Rubin.

Dato $\begin{cases} H_0 : \theta \leq \theta_0 \\ H_1 : \theta > \theta_0 \end{cases}$ Se T è stat suff per θ tale che la legge di T, ovvero $g(t, \theta)$, ha MLR

Allora $\forall t_0$ il test con $RC = \{T > t_0\}$ è UMP di livello $\alpha = \mathbb{P}_{\theta_0}(T > t_0)$

Oss. Ovvero quando ho l'ipotesi in questo modo e ho a disposizione una statistica suff con legge MLR

Dim.

$$\beta(\theta) = \mathbb{P}_\theta(T > t_0)$$

Mostriamo che T ha MLR $\implies$ la funzione potenza $\beta(\theta)$ è non decrescente in θ

Posto $\theta_1 < \theta_2$ e data la $F(t,\theta)$ la funzione di ripartizione di T

$$\frac{d}{dt}[F(t,\theta_1) - F(t,\theta_2)] = g(t,\theta_2) - g(t,\theta_1) = g(t,\theta_1)\left[\underbrace{\frac{g(t,\theta_2)}{g(t,\theta_1)}}_{\text{crescente in t}} - 1\right]$$

Questa derivata essendo crescente o è sempre positiva oppure passa da negativa a positiva

Quindi $[F(t,\theta_1) - F(t,\theta_2)]$ è o sempre crescente, oppure prima decresce e poi cresce

E quindi questa differenza raggiunge il massimo in $-\infty$ oppure in $+\infty$

Ma essendo funzioni di ripartizione vanno entrambe da 0 a 1 e quindi

$$\begin{cases} F(-\infty,\theta_2) - F(-\infty,\theta_1) = 0 \\ F(+\infty,\theta_2) - F(+\infty,\theta_1) = 0 \end{cases}$$

$$\implies \qquad F(t,\theta_2) - F(t,\theta_1) \leq 0 \iff F(t,\theta_2) \leq F(t,\theta_2)$$

$$\beta(\theta_1) = \mathbb{P}_{\theta_1} = 1 - F(t_0,\theta_1) \leq 1 - F(t_0,\theta_2) = \mathbb{P}_{\theta_2}(t > t_0) = \beta(\theta)$$

Oss. Definiamo X stocasticamente più grande di Y, $X \underset{st}{\geq} Y$, se $F_X(t) \leq F_Y(t)$

Esempi: 10.7 10.8

Quindi se T ha MLR allora $\beta(\theta)$ è crescente

$$\implies \sup_{\theta \leq \theta_0} \beta(\theta) = \beta(\theta_0) \implies \text{ il livello è } \alpha = \mathbb{P}_{\theta_0}(RC) = \mathbb{P}_{\theta_0}(T > t_0)$$

Partendo dal caso $\begin{cases} H_0 : \ \theta = \theta_0 \\ H_1 : \ \theta = \theta' \end{cases}$ con $\theta' > \theta_0$ quindi ho un'ipotesi semplice inclusa nel caso generale

$$\text{Allora } \ \widetilde{K} = \inf_{\tau = \{t > t_0\}} \frac{g(t,\theta')}{g(\theta_0)}$$

$$T > t_0 \iff \frac{g(t,\theta')}{g(\theta_0)} > \widetilde{K}$$

$$T > t_0 \iff \left\{ g(t,\theta') > \widetilde{k} g(t,\theta_0) \right\} \quad \text{è UMP per il corollario di N-P}$$

□

Si può dimostrare il teorema nel caso opposto: $\begin{cases} H_0 : \ \theta \geq \theta_0 \\ H_1 : \ \theta < \theta_0 \end{cases}$

In cui si ha UMP con $RC : \{T < t_0\}$ $\alpha = \mathbb{P}_{\theta_0}(T < t_0)$

Lezione 14 (28/03/23)

Definizione.

Un test è detto **Test Unione Intersezione**, o test UI

Se Θ_0 si può scrivere come un'intersezione di sottoinsiemi di Θ e se

$H_0 : \ \theta \in \Theta_0 = \bigcap_{\gamma \in \Gamma} \Theta_{0\gamma}$ VS $H_1 : \ \theta \in \Theta_0^C = \bigcup_{\gamma \in \Gamma} \Theta_{0\gamma}^C$

$\forall\, \gamma \quad H_{0\gamma} : \ \theta_0 \in \Theta_{0\gamma}$ VS $H_{1\gamma} : \ \theta \in \Theta_{0\gamma}^C$ con RC_γ

$\implies$ RC del test UI è $RC = \bigcup_{\gamma \in \Gamma} RC_\gamma$

Oss. Complementare dell'intersezione è l'unione dei complementari

Oss. La RC è unione perché devo accettare tutti gli $H_{0\gamma}$

Esempio: 14.1

Definizione.

Un test è detto **intersezione unione**, o test IU, quando si può scrivere

$$H_0 : \ \theta \in \Theta_0 = \bigcup_{\gamma \in \Gamma} \Theta_{0\gamma} \quad \text{VS} \quad H_1 : \ \theta \in \Theta_0^C = \bigcap_{\gamma \in \Gamma} \Theta_{0\gamma}^C$$

$$RC_{IU} = \bigcap_{\gamma \in \Gamma} RC_\gamma$$

Esempio: 14.2

Teorema.

Dato un test UI $\begin{cases} H_0 : \ \theta \in \bigcap\limits_{\gamma\in\gamma} \Theta_{0\gamma} \\ H_1 : \ \theta \in \bigcap\limits_{\gamma\in\gamma} \Theta_{0\gamma}^C \end{cases}$

Chiamiamo $\lambda_\gamma(\vec{x})$ la statistica del LRT per $H_{0\gamma} = \theta \in \Theta_{0\gamma}$

e $\lambda(\vec{x})$ la statistica del LRT per il test UI

Sia $T(\vec{x}) = \inf\limits_{\gamma} \lambda_\gamma(\vec{x})$ e siano

$$RC_T = \{\lambda_\gamma(\vec{x}) = c \quad \text{per qualche } \gamma\} = \{T(\vec{x}) \le c\}$$

$$RC_\lambda = \{\lambda(\vec{x}) \le c\}$$

Allora:

a) $T(\vec{x}) \ge \lambda(\vec{x}) \ \ \forall \vec{x}$

b) $\beta_T(\theta) \le \beta_\lambda(\theta) \ \ \forall \theta$

c) Se il LRT ha livello α, allora il test UI ha livello α

Dim.

a)

$$\Theta_0 = \bigcap_{\gamma\in\Gamma} \Theta_{0\gamma} \implies \Theta_0 \subset \Theta_{0\gamma} \ \forall\gamma$$

$$\lambda_\gamma(\vec{x}) = \frac{\sup_{\Theta_{0\gamma}} L(\theta, \vec{x})}{\sup_{\Theta} L(\theta, \vec{x})} \ge \frac{\sup_{\Theta_0} L(\theta, \vec{x})}{\sup_{\Theta} L(\theta, \vec{x})} = \lambda(\vec{x})$$

$$\implies T(\vec{x}) = \inf_{\gamma} \lambda_\gamma(\vec{x}) \ge \lambda(\vec{x})$$

b)

$$\beta_T(\theta) = \mathbb{P}_\theta(T(\vec{x}) \le c) \le \mathbb{P}_\theta(\lambda(\vec{x}) \le c) = \beta_\lambda(\theta)$$

c)

$$\text{Livello test UI } = \sup_{\theta\in\Theta_0} \beta_T(\theta) \le \sup_{\theta\in\Theta_0} \beta_\gamma(\theta) \le \alpha$$

□

Teorema.

Dato un test IU con $\Theta_0 = \bigcup\limits_{\gamma\in\Gamma} \Theta_{0\gamma}$

Sia α_γ il livello del test su $\Theta_{0\gamma}$ con RC_γ allora il test IU ha livello $\alpha = \sup\limits_{\gamma} \alpha_\gamma$

Dim.

$$\theta \in \Theta_0 \implies \theta \in \Theta_{0\gamma} \text{ per qualche } \gamma$$

$$\mathbb{P}_\theta(\vec{x} \in RC) \leq \mathbb{P}_\theta(\vec{x} \in RC_\gamma) \leq \alpha_\gamma \leq \sup_\gamma \alpha_\gamma$$

□

Vediamo questo teorema in un caso particolare del test IU, dove l'insieme dei $\Theta_{0\gamma}$ ha un numero di indici ben definito

Teorema.

Dato un test IU con $H_0: \ \theta \in \cup_{j=1}^k \Theta_{0j}$ e RC_j la RC del test $H_0 : \theta \in \Theta_{0j}$

Supponiamo che per qualche $i = 1...k \ \ \exists$ una successione di parametri $\theta_l \in \Theta_{0i}$ tale che

$$1) \lim_{l\to+\infty} \mathbb{P}_{\theta_l}(\vec{x} \in RC_i) = \alpha$$

$$2) \lim_{l\to\infty} \mathbb{P}_{\theta_l}(\vec{x} \in RC_j) = 1 \quad j = 1...k \ j \neq i$$

Allora il test IU con $RC = \bigcap_{j=1}^n RC_j$ ha livello α

Dim.

Dal teorema precedente RC ha livello α, vogliamo mostrare che ha proprio dimensione α

$$\theta_l \in \Theta_{0i} \in \Theta_0 = \underset{j}{\cup}\Theta_{0j}$$

$$\implies \sup_{\theta\in\Theta_0} \mathbb{P}_\theta(\vec{x} \in RC) \geq \lim_{l\to+\infty} \mathbb{P}_{\theta_l}(\vec{x} \in RC) = \lim_{l\to+\infty} \mathbb{P}_{\theta_l}(\vec{x} \in \underset{j}{\cap} RC_j) \geq$$

Vale per la disuguaglianza di Bonferroni $\mathbb{P}\left(\bigcap_{i=1}^n A_i\right) \geq \sum_{j=1}^n \mathbb{P}(A_i) - (n-1)$

$$\geq \lim_{t\to+\infty} \sum_{j=1}^k \mathbb{P}_{\theta_l}(\vec{x} \in RC_j) - (k-1) = (k-1)\cdot 1 + \alpha - (k-1) = \alpha$$

□

Esempio: 14.2

Oss. Definiamo i seguenti test: Unilateri o one-sided se $H_0: \ \theta \leq \theta_0 \quad H_1: \ \theta \geq \theta_0$

Bilateri o two-sided se $H_0: \ \theta = \theta_0 \quad H_1: \ \theta \neq \theta_0$

4.3 p-value

Definizione.

Il **p-value** $p(\vec{X})$ è una qualunque statistica tale che $0 \leq p(\vec{X}) \leq 1$

Voglio costruire i p-value di modo che valori piccoli di $p(\vec{X})$ siano a supporto di H_1

Definizione.

Una statistica p-value è **valida** se $\forall \theta \in \Theta_0$ e $\forall 0 \leq \alpha \leq 1 \quad \mathbb{P}_\theta(p(\vec{X}) \leq \alpha) \leq \alpha$

Questa definizione ha senso perché mi dice che sotto Θ_0 la probabilità di avere valori piccoli di p è piccola

Se p-value è valido allora lo posso usare per valutare dei test:

Posso costruire $RC = \{p(\vec{X}) \leq \alpha\}$ che ha livello α

$$\text{Perché} \quad \sup_{\theta \in \Theta_0} \mathbb{P}_\theta(RC) = \sup_{\theta \in \Theta_0} \mathbb{P}_\theta(p(\vec{X}) \leq \alpha) \leq \alpha \quad \forall \theta \in \Theta_0$$

Esempio: 14.3

Lezione 15 (03/04/23)

Teorema.

Supponiamo che sia $W(\vec{X})$ una statistica tale che valori grandi di W danno evidenza a favore di H_1

Sia $p(\vec{X}) = \sup_{\theta \in \Theta_0} \mathbb{P}_\theta(W(\vec{X}) \geq W(\vec{x}))$

Allora $p(\vec{X})$ è p-value valido

Esempio: 15.1

Oss. Uno statistico non usa il "trashold", ovvero non confronta il p-value per decidere se accettare o rifiutare il test

Nel senso che non dice "accetto se p-value > 0.05" e va a calcolare il valore del p-value. Infatti due valore di p-value 0.051 e 0.049 sono equivalenti per uno statistico

5 Stima Intervallare

5.1 Regioni di confidenze

Precedentemente avevamo costruito stimatori puntuali che portavano a stime puntuali

Adesso costruiamo stime intervallari che portano a intervalli di confidenza (i.e. $\theta \in \mathbb{R}$) o a regioni di confidenza (i.e. $\vec{\theta} \in \mathbb{R}^k$)

Definizione.

Una **stima intervallare** di un parametro reale θ è una coppia di statistiche $L(\vec{X}), U(\vec{X})$ tali che

$$L(\vec{x}) \leq U(\vec{x}) \ \forall \vec{x}$$

Allora la stima intervallare è l'intervallo $[L(\vec{X}); U(\vec{X})]$ e quindi la mia inferenza intervallare è

$$L(\vec{X}) \leq \theta \leq U(\vec{X})$$

Oss. Posso anche avere stime degeneri con $L = -\infty$ oppure $U = +\infty$

Ma tendenzialmente preferirò avere intervalli piccoli

Definizione.

Dato una stima intervallare $[L(\vec{X}; U(\vec{X}))]$ per θ

$\mathbb{P}_\theta \left(\theta \in [L(\vec{X}), U(\vec{X})]\right)$ è detta **probabilità di copertura**

Definizione.

Dato una stima intervallare $[L(\vec{X}; U(\vec{X}))]$ per θ

$\inf_\theta \mathbb{P}_\theta \left(\theta \in [L(\vec{X}), U(\vec{X})]\right)$ è detto **livello di confidenza**

Oss. Il livello di confidenza voglio sia alto, infatti lo chiameremo $(1 - \alpha)$ che voglio sia alto

Esempio: 15.2

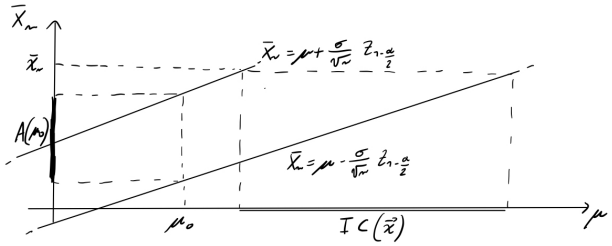
Esempio: 15.3

5.2 Metodi per trovare stimatori intervallari

(I) Inversione di un test d'ipotesi

Esempio: 15.4

$$\vec{X} \in A(\mu_0) \iff \mu_0 \in IC(\vec{X})$$



Test: Fissato parametro $= \mu_0$ individuo i dati campionari $(\vec{x})$ "consistenti" con tale valore $\implies$ accetto H_0

IC: Fissato il valore del campione individuo i valori del parametro che sono "compatibili"

Teorema.

1) $\forall \theta_0 \in \Theta$ sia $A(\theta_0)$ la regione di accettazione di un test di livello α per $H_0 : \theta = \theta_0$

$\implies \forall \vec{x}$ sia $C(\vec{x}) = \{\theta_0 : \vec{x} \in A(\theta_0)\}$

Allora $C(\vec{x})$ è una regione di confidenza di livello $1 - \alpha$ per θ

2) Viceversa, sia $C(\vec{X})$ una regione di confidenza per θ di livello $1 - \alpha$

$\implies \theta_0 \in \Theta$ sia $A(\theta_0) = \{\vec{X} : \theta_0 \in C(\vec{X})\}$

Allora $A(\vec{\theta_0})$ è regione di accettazione di un test di livello α per $H_0 : \theta = \theta_0$

Dim.

1) $\mathbb{P}_{\theta_0}(\theta \in C(\vec{x})) = \mathbb{P}_{\theta_0}(\vec{x} \in A(\theta_0)) \geq 1 - \alpha$

2) $\mathbb{P}_{\theta_0}(\vec{X} \notin A(\theta_0)) = \mathbb{P}_{\theta_0}(\theta \notin C(\vec{X})) \leq \alpha$

□

Lezione 16 (04/04/23)

Esempio: 16.1

(II) Metodo basato sulla quantità pivotale

Definizione: Pivot.

Una variabile aleatoria $Q(\vec{X}, \vec{\theta}) = Q(X_1, ..., X_n, \vec{\theta})$ è detta **quantità pivotale**, o pivot, se la sua legge $\mathcal{L}(Q)$ non dipende da $\vec{\theta}$

Oss. Dipende anche dai parametri e quindi non è una statistica

Esempio: 16.2, 16.3, 16.4

Data una quantità pivotale Q e fissata $\alpha \in [0, 1]$

Possiamo sempre trovare una coppia $[a, b]$ che non dipende da θ tc

$$\mathbb{P}(a \leq Q \leq b) \geq 1 - \alpha \quad \Longrightarrow \quad C(\vec{X}) = \{\theta : a \leq Q \leq b\} = \text{ regione di confidenza di livello } 1 - \alpha$$

L'obbiettivo è trovare l'intervallo di confidenza in funzione di θ di livello $1 - \alpha$, ovvero vogliamo invertire Q:

$$a \leq Q(\vec{X}, \theta) \leq b \Longleftrightarrow h(\vec{X}, a) \leq \theta \leq g(\vec{X}, b)$$

Esempio: 16.5, 16.6

Definizione.

Una legge $f_X(x)$ è detta **unimodale** se $\exists x^*$ tale che $f_X(x)$ è non decrescente $\forall x \leq x^*$ ed è non crescente $\forall x > x^*$

Oss. Ovvero se esiste unico un punto di massimo

Teorema.

Sia $f_X(x)$ una densità unimodale. Se l'intervallo $[a, b]$ soddisfa:

(1) $\int_a^b f_X(x)\, dx = 1 - \alpha$

(2) $f_X(a) = f_X(b) > 0$

(3) $a \leq x^* \leq b$

Allora $[a, b]$ è l'intervallo di lunghezza minima tra tutti quelli che soddisfano la (1)

Dim. (*)

Sia $[a', b']$ un intervallo con $(b' - a') < (b - a)$, voglio mostrare che non verifica la proprietà (1)

Fissiamo $a' \leq a$ avremo più casi:

i) $b' \leq a \quad a' \leq b' \leq a \leq x^* \quad$ quindi sono nella parte crescente della densità unimodale e quindi l'integrale tra a' e b' è più piccolo del rettangolo con altezza $f_x(b')$

$$\implies \int_{a'}^{b'} f_X(x)\,dx \leq f_X(b')(b' - a') \leq f_X(a)(b' - a') < f_X(a)(b - a) \leq \int_a^b f_X(x)\,dx = 1 - \alpha$$

ii) $b' > a \quad a' \leq a < b'b$

$$\implies \int_{a'}^{b'} f_X(x)\,dx = \int_a^b f_X(x)\,dx + \int_{a'}^{a} f_X(x)\,dx - \int_{b'}^{b} f_X(x)\,dx = (1-\alpha) + \left[\int_{a'}^{a} f_X(x)\,dx - \int_{b'}^{b} f_X(x)\,dx\right]$$

Voglio mostrare che la parte dentro le quadre sia minore di zero

$$\text{Essendo sulla parte crescente} \quad \int_{a'}^{a} f_X(x)\,dx \leq f_X(a)(a - a')$$

$$\text{Essendo sulla parte decrescente} \quad \int_{b'}^{b} f_X(x)\,dx \geq f_X(b)(b - b')$$

$$\implies \left[\int_{a'}^{a} f_X(x)\,dx - \int_{b'}^{b} f_X(x)\,dx\right] \leq f_X(a)(a - a') - f_X(b)(b - b') \overset{(2)}{=} f_X(a)[(b' - a') - (b - a)] < 0$$

□

Applicazione del teorema:

Dato Q unimodale, allora l'intervallo a,b $\mathbb{P}(a \leq Q \leq b) = 1 - \alpha$ ha lunghezza minima se scelgo $f_Q(b) = f_Q(a)$ e quindi $a \leq q^* \leq b$

$$IC = [h(\vec{X}, a, b) \leq \theta \leq g(\vec{X}, a, b)]$$

Se lunghezza di IC è proporzionale a $(b - a)$ allora uso il teorema (vedi esempio della gaussiana)

Se invece la lunghezza di IC non è proporzionale a $(b - a)$ allora il teorema non serve

Esempio: 16.7

6 Teoria asintotica

Lezione 17 (12/04/23)

Studio delle proprietà di stimatori, test ... quando l'ampiezza del campione: $n \to +\infty$

> **Definizione.**
> Una successione di stimatori W_n è **consistente** se $\forall \varepsilon > 0 \ \lim_{n\to+\infty} \mathbb{P}(|W_n - \theta| < \varepsilon) = 1$
> ovvero se $W_n \underset{\mathbb{P}}{\to} \theta$

Esempio:
Media campionaria $\overline{X}_n$: X_1 , $\frac{X_1+X_2}{2}$, $\frac{X_1+X_2+X_3}{3}$ ...
$\overline{X}_n$ è consistente per $\mathbb{E}[X_i] = \mu$

Sappiamo che $Y_n \xrightarrow{qc} Y \implies Y_n \xrightarrow{\mathbb{P}} Y \implies Y_n \xrightarrow{\mathcal{L}} Y$
$Y_n \xrightarrow{\mathcal{L}} Y \not\Longrightarrow Y_n \xrightarrow{\mathbb{P}} Y \not\Longrightarrow Y_n \xrightarrow{qc} Y$
$Y_n \xrightarrow{\mathcal{L}} c \implies Y_n \xrightarrow{\mathbb{P}} Y$

Possiamo valutare la consistenza usando Chebyshev:

$$\mathbb{P}(|W_n - \theta| \geq \varepsilon) \leq \frac{\mathbb{E}\left[|W_n - \theta|^2\right]}{\varepsilon^2} = \frac{MSE(W_n)}{\varepsilon^2}$$

Quindi se $MSE(W_n) \to 0$ allora W_n è consistente per θ e $W_n \xrightarrow{\mathbb{P}} \theta$

> **Definizione.**
> Sia W_n una successione di stimatori tali che
>
> $$\sqrt{n}\left[W_n - \tau(\theta)\right] \xrightarrow{\mathcal{L}} \mathcal{N}(0, \sigma^2)$$
>
> Allora W_n è asintoticamente gaussiano e σ^2 è detta **varianza asintotica**

Questo risultato vale per il TCL, che dice:

$$\sqrt{n}\left(\overline{X}_n - \mathbb{E}[X_i]\right) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \mathrm{Var}[X_i])$$

Quindi $\overline{X}_n \approx \mathcal{N}\left(\mathbb{E}[X_i], \frac{\mathrm{Var}[X_i]}{n}\right)$
Oss. In questo caso si parla di asintotica normalità (AN), attenzione che quando si usa il limite non possiamo

tenere il valore n come parametro

Mostriamo che la varianza asintotica è diversa dal limite delle varianze degli W_n

Esempio: 17.1

Corollario: Asintotica normalità $\implies$ consistenza

$$\sqrt{n}\,(W_n - \tau(\theta)) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \sigma^2) \implies W_n \xrightarrow{\mathbb{P}} \tau(\theta)$$

Dim.

$$(W_n - \tau(\theta)) = \frac{\sigma}{\sqrt{n}} \left(\frac{W_n - \tau(\theta)}{\sigma} \right) \cdot \sqrt{n} \xrightarrow{\mathcal{L}} 0 \quad \text{per il teorema di Slutsky}$$

$$\text{Infatti} \quad \frac{\sigma}{\sqrt{n}} \xrightarrow{q.c.} 0 \qquad \frac{W_n - \tau(\theta)}{\sigma} \cdot \sqrt{n} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1)$$

$$(W_n - \tau(\theta)) \xrightarrow{\mathcal{L}} 0 \implies (W_n - \tau(\theta)) \xrightarrow{\mathbb{P}} 0 \quad \text{perché il limite è una costante}$$

□

Definizione.

Successione di stimatori W_n è **asintoticamente efficiente** per $\tau(\theta)$ se

$$\sqrt{n}\,[W_n - \tau(\theta)] \xrightarrow{\mathcal{L}} \mathcal{N}(0, v(\theta)) \quad \text{con} \quad v(\theta) = \frac{[\tau'(\theta)]^2}{\mathbb{E}\left[\left(\frac{\partial}{\partial \theta} log\, f(x,\theta) \right)^2 \right]} = \frac{[\tau'(\theta)]^2}{I_1(\theta)}$$

$$\equiv W_n \approx \mathcal{N}\left(\tau(\theta), \frac{[\tau'(\theta)]^2}{I_n(\theta)} \right)$$

Teorema: Efficienza asintotica degli MLE.

Siano $X_1, ..., X_n \sim f(x,\theta)$ con $f(x,\theta)$ che soddisfa le ipotesi di regolarità di C.R. e $\widehat{\theta}_{MLE}$ stimatore ML per θ

$$\implies \sqrt{n}\left[\tau(\widehat{\theta}_{MLE}) - \tau(\theta) \right] \xrightarrow{\mathcal{L}} \mathcal{N}(0, v(\theta)) \quad \text{con} \quad v(\theta) = \frac{[\tau'(\theta)]^2}{I_1(\theta)}$$

Ovvero gli MLE sono asintoticamente efficienti

Oss. Nel caso in cui non valgano le ipotesi di CR, dovrò lavorare a mano, vediamo il seguente:

Esempio: 17.2

Definizione.

Siano W_n e V_n due successioni di stimatori tali che

$$\sqrt{n}\left[W_n - \tau(\theta) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \sigma_W^2)\right] \qquad \sqrt{n}\left[V_n - \tau(\theta) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \sigma_V^2)\right]$$

Allora si chiama **ARE** o asyntotic relative efficiency di V_n rispetto a W_n

$$ARE(V_n; W_n) = \frac{\sigma_W^2}{\sigma_V^2}$$

Oss. Questa frazione ci dice quale stimatore è più efficiente

Esempio: 17.3

Teorema: Metodo delta 1.

Siano Y_n tale che $\sqrt{n}(Y_n - \theta) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \sigma^2)$ e g una funzione tale che $\exists\, g'(\theta) \neq 0$

$$\text{Allora} \quad \sqrt{n}\,(g(Y_n) - g(\theta)) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \sigma^2(g'(\theta))^2)$$

Teorema: Metodo delta 2.

Siano Y_n tale che $\sqrt{n}(Y_n - \theta) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \sigma^2)$ e g una funzione tale che $\exists\, g'(\theta) = 0\ \exists\, g''(\theta) \neq 0$

$$\text{Allora} \quad n\,(g(Y_n) - g(\theta)) \xrightarrow{\mathcal{L}} \frac{\sigma^2}{2} g''(\theta) \cdot \chi_1^2$$

Esempio: 17.4

Lezione 18 (17/04/23)

Definizione.

L'**odds** è il rapporto tra la probabilità che un evento accada fratto la probabilità che non accada, per una Bernulli è $\dfrac{p}{1-p}$ invece l'**odds ratio** è il rapporto tra odds: $\dfrac{\frac{p_1}{1-p_1}}{\frac{p_2}{1-p_2}}$

Un valore molto importante per le Bernulli è il $logit(p) = log\left(\dfrac{p}{1-p}\right)$

$logit(p) : [0,1] \to \mathbb{R}$ questa funzione è monotona crescente e quindi invertibile infatti

$$y = logit(p) = log\left(\frac{p}{1-p}\right) \iff e^y = \frac{p}{1-p} \iff e^y = p(1+e^y) \iff p = \frac{e^y}{1+e^y}$$

Esempio: 18.1

Teorema.

Dato un test $\begin{cases} H_0 : \theta = \theta_0 \\ H_1 : \theta \neq \theta_0 \end{cases}$

Siano $X_1, ..., X_n \sim f(x, \theta)$ e $\widehat{\theta}_{MLE}$ stimatore di massima verosimiglianza per θ

Allora sotto l'ipotesi $H_0 \quad -2\,log(\lambda(\vec{x})) \xrightarrow{\mathcal{L}} \chi^2_1$

Oss. Servono le condizioni di regolarità (famiglia esponenziale)

Dim.

$$-2\,log(\lambda(\vec{x})) = -2\,log\left(\frac{L(\theta_0, \vec{x})}{L(\widehat{\theta}_{MLE}, \vec{x})}\right) = -2\,l(\theta_0, \vec{x}) + 2l(\widehat{\theta}_{MLE}, \vec{x})$$

Sviluppiamo il primo termine con Taylor (dovrò andare al secondo ordine perché derivata prima nulla)

$$l(\theta_0, \vec{x}) = l(\widehat{\theta}_{MLE}, \vec{x}) + l'(\widehat{\theta}_{MLE}, \vec{x})(\widehat{\theta}_{MLE} - \theta_0) + \frac{1}{2}l''(\widehat{\theta}_{MLE}, \vec{x})(\widehat{\theta}_{MLE} - \theta_0)^2 + ...$$

Però dato che $\widehat{\theta}_{MLE}$ è il massimo ho $(\widehat{\theta}_{MLE}, \vec{x}) = 0$

$$-2\,log(\lambda(\vec{x})) = -2\,l(\widehat{\theta}_{MLE}, \vec{x}) - l''(\widehat{\theta}_{MLE}, \vec{x})(\widehat{\theta}_{MLE} - \theta_0)^2 + 2l(\widehat{\theta}_{MLE}, \vec{x}) = -l''(\widehat{\theta}_{MLE}, \vec{x})(\widehat{\theta}_{MLE} - \theta_0)^2$$

Si può dimostrare che

$$I_n(\widehat{\theta}_{MLE}) = -l''(\widehat{\theta}_{MLE}, \vec{x}) \qquad \frac{I_n(\widehat{\theta}_{MLE})}{n} \xrightarrow[\mathbb{P}]{q.c.} I(\theta)$$

Sotto H_0 ho

$$-2\,log(\lambda(\vec{x})) = -\underbrace{\frac{l''(\widehat{\theta}_{MLE}, \to)}{n} \cdot \frac{1}{I(\theta_0)}}_{\to 1} \cdot \underbrace{\left(\frac{\sqrt{n}(\widehat{\theta}_{MLE} - \theta_0)}{\frac{1}{\sqrt{I(\theta_0)}}}\right)^2}_{\xrightarrow{\mathcal{L}} \chi^2_1}$$

Per Slutsky $\quad -2\,log(\lambda(\vec{x})) \xrightarrow{\mathcal{L}} \chi^2_1$

□

7 Limiti della statistica parametrica

Fin'ora abbiamo visto la statistica parametrica, per cui dato un campione $X_1, ..., X_n$ cercavamo la legge del campione tra le leggi del tipo $f(x, \vec{\theta})$, per cui ci riducevamo dalla ricerca dallo spazio infinito dimensionale delle leggi, alla ricerca finito dimensionale dei parametri: $\vec{\theta} \in \Theta \subseteq \mathbb{R}^k$

Però quest'ipotesi, che è verificata in molte situazioni, non è sempre verificata, in questi casi possiamo procedere in questo modo:

1) Esistono tecniche di inferenza per dati generati da leggi non parametriche

2) Metodi per controllare se i dati vengono generati da leggi parametriche, detti test di buon adattamento:

Nell'ambito della regressione lineare assumeremo che i dati siano gaussiani, che metodi useremo per **verificare la gaussianità**?

Possiamo guardare la distribuzione dei dati, è necessario fare una **rappresentazione grafica** dei dati, per esempio con un istogramma, da cui è possibile notare subito se la distribuzione assomiglia a una gaussiana

Poi si fanno i **qqplot** ovvero un grafico con in ascissa i quantili teorici Z_α e in ordinata i quantili empirici χ_α

Sappiamo che i quantili di una gaussiana non standard hanno forma del tipo: $\chi_\alpha = \mu + \sigma Z_\alpha$
Quindi verifico che il qqplot abbia un andamento lineare, o quasi

Infine posso usare dei test non parametrici del tipo: $\begin{cases} H_0 : F_X \sim \mathcal{N}(\mu, \sigma^2) \\ H_1 : F_X \not\sim \mathcal{N}(\mu, \sigma^2) \end{cases}$
Un esempio è lo **Shapiro test** che faremo con in laboratorio con il software

8 Modelli di regressione

Ci sono modelli di regressione lineari, lineari generalizzati, non lineari ...

Per ogni unità statistica i $(y_i, z_{i1}, ..., z_{ip})$
Osservo y_i dato generato dalla VA Y dipendente e $z_{i1}, ..., z_{ip}$ dati generati dalle VA Z_j indipendenti, covariate, feature

Modellare la Y (ovvero la legge che genera $y_1, ..., y_n$) in funzione dei valori delle covariate

In generale ci sarà una relazione $t_i = g(z_{i1}, ..., z_{ip})$

Nel caso dei modelli lineari assumiamo $Y = \beta_0 + \beta_1 Z_1 + ... + \beta_p Z_p + \varepsilon$
Dove β_j sono i parametri incogniti, Z_j sono deterministiche e ε è l'errore aleatorio

Stimeremo i β_j in funzione di Y e ε per due motivi: spiegare e prevedere

Lezione 19 (26/04/23)

Osserveremo un set di dati $(y_i, z_{i1}, ..., z_{ir})$ e dovremo capire se questo è stato generato dalle variabili Y, Z_i del tipo $Y = \beta_0 + \beta_1 Z_1 + ... + \beta_r Z_r + \varepsilon$

Vorremo riuscire a prevedere Y = variabile dipendente attraverso le $Z_1 ... Z_r$ variabili indipendenti

Mentre ε è una VA tc $\mathbb{E}[\varepsilon] = 0$ e $\text{Var}[\varepsilon] = \sigma^2$
σ^2 sarà il parametro incognito da stimare

$$\mathbb{E}[Y|Z_1 ... Z_r] = \beta_0 + \beta_1 Z_1 + ... \beta_r Z_r$$

Abbiamo linearità rispetto ai parametri β_i quindi potremmo stimare, per esempio, con $Z_2 = Z_1^2$, ma non con $Y = sen(\beta_0 + \beta_1 Z_1) + \varepsilon$

L'obbiettivo è quello di stimare i parametri incogniti $(\beta_0, ..., \beta_r, \sigma^2)$ per diversi motivi:

- Spiegare e interpretare la relazione tra Y e Z_i

- Prevedere il valore di Y in corrispondenza di Z_i (non ancora osservati)

Useremo i GOF = indici di "goodness of fit" per valutare la bontà del modello

Supponiamo invece di avere più dati:

$$\forall i = 1...n \quad (y_i, z_{i1}, ..., z_{i\varepsilon}) \quad \varepsilon_i \; iid \; \mathbb{E}[\varepsilon_i] = 0 \; \mathrm{Var}[\varepsilon_i] = \sigma^2$$

$$\vec{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \quad \mathbb{Z} = \begin{bmatrix} 1 & z_{11} & & z_{1r} \\ \vdots & \vdots & \dots & \vdots \\ 1 & z_{n1} & & z_{nr} \end{bmatrix} \quad \vec{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix} \quad \vec{\beta} = \begin{bmatrix} \beta_0 \\ \vdots \\ \beta_r \end{bmatrix}$$

Possiamo riscriveremo il probelma nel modo seguente:

$$\vec{Y} = \mathbb{Z}\vec{\beta} + \vec{\varepsilon}$$

Ovverp $y_i = \sum_{j=0}^{r} \beta_i z_{1j} \varepsilon_i = \beta_0 + \beta_1 z_{i1} + ... + \beta_r z_{ir} + \varepsilon_i$
Oss. Di solito avremo $r << n$

Con $r = 1$ si dirà regressione lineare semplice $\quad y = \beta_0 + \beta_1 z_i + \varepsilon_i$

Mentre con $r > 1$ avremo regressione lineare multipla

Possiamo dividere ulteriormente i casi in base alle Z che possono essere continue, categoriche o miste

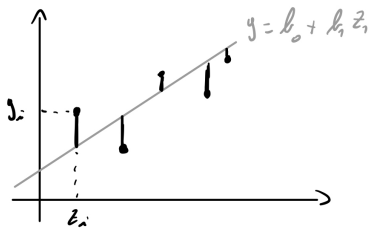
Se non facessimo ipotesi sulla legge di ε, avremmo la distribuzione di $\sum \beta_i Z_i$, ma non avendo la legge di ε non possiamo stimare usando la Likelihood

Per stimare $\vec{\beta}, \sigma^2$ useremo il OLS, Ordinary last squared, ovvero stima ai minimi quadrati

$$\vec{y} \in \mathbb{R}^n \quad \widehat{\vec{\beta}}_{LS} = \underset{\vec{b} \in \mathbb{R}^{r+1}}{Argmin} (\vec{y} - \mathbb{Z}\vec{b})^T (\vec{y} - \mathbb{Z}\vec{b})$$

Dove $(\vec{y} - \mathbb{Z}\vec{b})^T (\vec{y} - \mathbb{Z}\vec{b}) = \sum_{i=1}^{n} (y_i - (\mathbb{Z}\vec{b})_i)^2 = \sum_{i=1}^{n} (y_i - \sum_j b_j z_{ij})^2$

Nel caso di $r = 1$, avremo che b_0, b_1 saranno i coefficienti della retta che minimizza i quadrati delle distanze dai dati (z_i), ovvero $\sum_{i=1}^{n} (y_i - (b_0 + b_1 z_i))^2$



Teorema.

Supponiamo che $\mathbb{Z}$, detta matrice disegno, abbia rango $r+1$

1) $\widehat{\vec{\beta}}_{LS} = (\mathbb{Z}^T\mathbb{Z})^{-1}\mathbb{Z}^T\vec{y}$ sono stimatori lineari di $\vec{y}$

2) Sia $H = \mathbb{Z}(\mathbb{Z}^T\mathbb{Z})^{-1}\mathbb{Z}^T$ la matrice "cappuccio" (o "hat matrix")

$$\widehat{\vec{y}} = H\vec{y} = \mathbb{Z}\widehat{\vec{\beta}}_{LS} \quad \widehat{\vec{y}} = \text{ "fitted values"}$$

$$\widehat{\vec{\varepsilon}} = (\vec{y} - \widehat{\vec{y}}) = (\mathbb{1} - H)\vec{y} \quad \widehat{\vec{\varepsilon}} = \text{ "residual values" = residui}$$

Inoltre si dimostra che $\mathbb{Z}^T\widehat{\vec{\varepsilon}} = 0$ ovvero $\widehat{\vec{y}}^T\widehat{\vec{\varepsilon}} = 0$ ovvero $\widehat{\vec{y}} \perp \widehat{\vec{\varepsilon}}$

Vediamo l'idea grafica di questo teorema

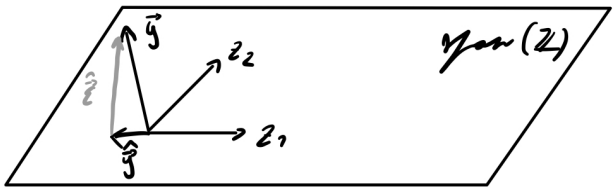


Figure 1: $\widehat{y}$ è la proiezione ortogonale di y sul sottospazio lineare generato dalle colonne di $\mathbb{Z}$, mentre $\widehat{\varepsilon}$ è la differenza tra i due vettori y

Dim. (*)

Per costruire la proiezione devo trovare una base ortonormale per $Span(\mathbb{Z})$

Considero $\mathbb{Z}^T\mathbb{Z}$ forma quadratica definita positiva $\quad \mathbb{Z}^T\mathbb{Z} = \sum_{i=1}^{r+1} \lambda_i \vec{e}_i \vec{e}_i^T$ con $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_{r+1} > 0$ i λ_i sono gli autovalori, mentre gli $\vec{e}_i$ sono autovettori

Quindi $(\mathbb{Z}^T\mathbb{Z})^{-1} = \sum_{i=1}^{r+1} \frac{1}{\lambda_i} \vec{e}_i \vec{e}_i^T$

Definisco $\vec{q}_i = \frac{1}{\sqrt{\lambda_i}} \mathbb{Z}\vec{e}_i \quad i = 1, ..., r+1$ e per costruzione $\vec{q}_i \in Span(\mathbb{Z})$

$$\vec{q}_i^T\vec{q}_j = \frac{1}{\sqrt{\lambda_i\lambda_j}} \vec{e}_i^T(\mathbb{Z}^T\mathbb{Z})\vec{e}_j = \frac{1}{\sqrt{\lambda_i\lambda_j}} \vec{e}_i^T\lambda_j\vec{e}_j = \sqrt{\frac{\lambda_j}{\lambda_i}}\vec{e}_i^T\vec{e}_j = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases}$$

$$\implies \quad \vec{q}_i \text{ è base ortonormale di } Span(\mathbb{Z})$$

Per concludere scrivo

$$\vec{y} = \sum_{i=1}^{r+1} \vec{q}_i(\vec{q}_i^T\vec{y}) = \sum_{i=1}^{r+1} \frac{1}{\lambda_i} \mathbb{Z}\vec{e}_i(\mathbb{Z}\vec{e}_i)\vec{y} = \sum_{i=1}^{r+1} \frac{1}{\lambda_i} \mathbb{Z}\vec{e}_i\vec{e}_i^T\mathbb{Z}^T\vec{y} =$$

$$= \mathbb{Z}\left(\sum_{i=1}^{r+1} \frac{1}{\lambda_i} \vec{e}_i \vec{e}_i^T\right) \mathbb{Z}^T \vec{y} = \underbrace{\mathbb{Z}(\mathbb{Z}^T\mathbb{Z})^{-1}}_{H} \vec{y} = \mathbb{Z}\hat{\vec{\beta}}_{LS}$$

H è un proiettore ortogonale su un sottospazio, quindi valgono: $H^T = H \quad H^2 = H$ quindi

$$\hat{\vec{\varepsilon}} = (\vec{y} - \hat{\vec{y}}) = (\mathbb{1} - H)\vec{y}$$

$$\mathbb{Z}^T \hat{\vec{\varepsilon}} = \mathbb{Z}^T(\mathbb{1} - \mathbb{Z}(\mathbb{Z}^T\mathbb{Z})^{-1}\mathbb{Z}^T)\vec{y} = \mathbb{Z}^T\vec{y} - \mathbb{Z}^T\mathbb{Z}(\mathbb{Z}^T\mathbb{Z})^{-1}\mathbb{Z}^T\vec{y} = 0$$

□

Se abbiamo $rango(\mathbb{Z}) = k < r + 1$, allora

$$\mathbb{Z}^T\mathbb{Z} = \sum_{i=1}^{r+1} \lambda_i \vec{e}_I \vec{e}_i^T \quad \text{dove } \lambda_1 \geq \lambda_2 \geq ... \geq \lambda_k \quad \lambda_{k+1} = ... = \lambda_{r+1} = 0$$

Non possiamo fare l'inversa di questa matrice ~~$(\mathbb{Z}^T\mathbb{Z})^{-1}$~~

Quindi useremo la pseudo inversa o inversa generalizzata $(\mathbb{Z}^T\mathbb{Z})^-$ e varrà tutto allo stesso modo

Oss. Seguento quest'idea, vedremo che avremo più problemi nel caso in cui le covariate hanno delle collinearità, ovvero se $\mathbb{Z}$ ha colonne dipendenti

Decomposizione della varianza

Sappiamo che:

$$\vec{y} = \hat{\vec{y}} + \hat{\vec{\varepsilon}} \quad \text{con } \hat{\vec{y}} \perp \hat{\vec{\varepsilon}}$$

Quindi posso applicare il teorema di Pitagora in $\mathbb{R}^n$

$$\vec{y}^T\vec{y} = \hat{\vec{y}}^T\hat{\vec{y}} + \hat{\vec{\varepsilon}}^T\hat{\vec{\varepsilon}}$$

$$\sum_{i=1}^n y_i^2 = \sum_{i=1}^n \hat{y}_i^2 + \sum_{i=1}^n \hat{\varepsilon}_i^2 \iff \sum_{i=1}^n y_i^2 - n\overline{y}^2 = \sum_{i=1}^n \hat{y}_i^2 + \sum_{i=1}^n \hat{\varepsilon}_i^2 - n\overline{y}^2$$

Osservazioni (dove $\overline{y}$ è la media):

- $\sum_{i=1}^n (y_i - \overline{y}^2) = \sum y_i^2 - n\overline{y}^2$
- $\overline{y} = \overline{\hat{y}}$ ovvero hanno la stessa media

Dove $n\overline{\hat{y}} = \sum_{i=1}^n \hat{y}_i = \hat{\vec{y}}^T\vec{\mathbf{1}} = [\mathbb{Z}(\mathbb{Z}^T\mathbb{Z})\mathbb{Z}^T\hat{y}]^T\vec{\mathbf{1}} = \vec{y}^T[H]\vec{\mathbf{1}} = \vec{y}^T\vec{\mathbf{1}}$ se $\vec{\mathbf{1}} = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} \in \mathbb{Z}$

$$\sum_{i=1}^n (y_i - \overline{y})^2 = \sum_{i=1}^n (\hat{y}_i - \overline{\hat{y}})^2 + \sum_{i=1}^n \hat{\varepsilon}_i^2$$

Formula della decomposizione della varianza

Questa ci dice che la variabilità (somma degli scarti dalla media al quadrato) la posso decomporre in variabilità dei fittati più l'errore

Lezione 20 (02/05/23)

$$\underbrace{\sum_{i=1}^{n}(y_i-\overline{y})^2}_{SS_{tot}}=\underbrace{\sum_{i=1}^{n}(\widehat{y}_i-\overline{\widehat{y}})^2}_{SS_{reg}}+\underbrace{\sum_{i=1}^{n}\widehat{\varepsilon}_i^2}_{SS_{res}}$$

Definiamo:

SS_{tot} la somma totale degli scarti al quadrato della media

SS_{reg} la somma dei quadrati degli scarti previsti

SS_{res} somma dei quadrati dei residui

Definizione.

Sia il coefficiente di determinazione $R^2=\dfrac{SS_{reg}}{SS_{tot}}=\dfrac{\sum_{i=1}^{n}(\widehat{y}_i-\overline{\widehat{y}})^2}{\sum_{i=1}^{n}(y_i-\overline{y})^2}=1-\dfrac{SS_{reg}}{SS_{tot}}=1-\dfrac{\sum_{i=1}^{n}\widehat{\xi}_i^2}{\sum_{i=1}^{n}(y_i-\overline{y})^2}$

Sappiamo che $0\le R^2\le 1$

Inoltre R^2 quantifica la proporzione di variabilità dei dati spiegata dal modello di regressione

Oss. Se $R^2=1$ abbiamo $\sum_{i=1}^{n}(y_i-\overline{y})^2=0$, questo implica che abbiamo interpolazione perfetta, ma questo non ci piace perché non abbiamo più l'indice che indica la variabilità dei dati

Se invece $R^2=0$ cioè $\sum_{i=1}^{n}(\widehat{y}_i-\overline{\widehat{y}})^2 \quad \forall i \widehat{y}_i=\overline{y}$ e quindi i proiettori non influenzano la risposta

8.1 Regressione lineare semplice

$L=\sum_{i=1}^{n}(y_i-\beta_0-\beta_1 Z_i)^2$ vogliamo trovarne il massimo

$$\begin{cases}\dfrac{\partial L}{\partial\beta_0}=0\\ \dfrac{\partial L}{\partial\beta_1}=0\end{cases}\iff\begin{cases}-2\sum_{i=1}^{n}(y_i-\beta_0-\beta_1Z_i)=0\\ -2\sum_{i=1}^{n}(y_i-\beta_0-\beta_1Z_i)\cdot z_i=0\end{cases}\iff\begin{cases}n\beta_0+\beta_1\sum z_i=\sum y_i\\ \beta_0\sum z_i+\beta_1\sum z_i^2=\sum y_iz_i\end{cases}$$

$$\iff\begin{cases}\beta_0=\dfrac{\sum y_i}{n}-\beta_1\dfrac{\sum z_i}{n}\\ \dfrac{\sum y_i}{n}\sum z_i-\beta_1\dfrac{(\sum z_i)^2}{n}+\beta_1\sum z_i^2=\sum y_iz_i\end{cases}\iff\begin{cases}\widehat{\beta}_0=\overline{y}-\widehat{\beta_1}\overline{z}\\ \widehat{\beta}_1=\dfrac{\sum(y_i-\overline{y})(z_i-\overline{z})}{\sum(z_i-\overline{z})^2}\end{cases}$$

Oss. Questo equivale a risolvere $\widehat{\vec{\beta}} = (\mathbb{Z}^T\mathbb{Z})^{-1}\mathbb{Z}^T\vec{y}$

Ma nella forma trovata è più interpretabile, infatti possiamo vedere la prima equazione come il fatto che il "baricentro" $(\overline{y}, \overline{z})$ appartenga alla retta di fitting

Mentre $\widehat{\beta}_1$, ovvero la pendenza della retta, è la correlazione tra y e z

Teorema.

Dati gli stimatori ai minimi quadrati $\vec{y} = \mathbb{Z}\vec{\beta} + \vec{\varepsilon}$ $\widehat{\vec{\beta}}_{LS} = (\mathbb{Z}^T\mathbb{Z})^{-1}\mathbb{Z}^T\vec{y}$

1) $\mathbb{E}\left[\widehat{\vec{\beta}}_{LS}\right] = \vec{\beta}$ sono non distorti per $\vec{\beta}$

2) $Cov\left[\widehat{\vec{\beta}}_{LS}\right] = \sigma^2(\mathbb{Z}^T\mathbb{Z})^{-1}$

3) $\mathbb{E}\left[\widehat{\vec{\varepsilon}}\right] = 0$

4) $Cov\left[\widehat{\vec{\varepsilon}}\right] = \sigma^2(1-H)$

5) $\mathbb{E}\left[\widehat{\vec{\varepsilon}}^T\widehat{\vec{\varepsilon}}\right] = \sigma^2(n-r-1)$ dove Z è n x r-1 r= numero di intercette

Da quest'ultimo risultato otteniamo che $S^2 = \dfrac{\widehat{\vec{\varepsilon}}^T\widehat{\vec{\varepsilon}}}{n-r-1} = \dfrac{\sum \widehat{\varepsilon}_i^2}{n-r-1}$ è stimatore non distorto per σ^2

Dim.

1)

$$\mathbb{E}\left[\widehat{\vec{\beta}}_{LS}\right] = \mathbb{E}\left[(\mathbb{Z}^T\mathbb{Z})^{-1}\mathbb{Z}^T\vec{y}\right] = (\mathbb{Z}^T\mathbb{Z})^{-1}\mathbb{Z}^T\mathbb{E}\left[\vec{y}\right] = (\mathbb{Z}^T\mathbb{Z})^{-1}\mathbb{Z}^T\mathbb{E}\left[\mathbb{Z}\vec{\beta} + \vec{\varepsilon}\right] = (\mathbb{Z}^T\mathbb{Z})^{-1}\mathbb{Z}^T\mathbb{Z}\vec{\beta} = \vec{\beta}$$

2)Sappiamo che $Cov(A\vec{y}) = A\,Cov(\vec{y})A^T$

E che $Cov(\vec{y}) = Cov(\mathbb{Z}\vec{\beta} + \vec{\varepsilon}) = Cov(\vec{\varepsilon}) = \sigma^2\vec{1}_n$ perché $\mathbb{Z}\vec{\beta}$ non è aleatorio, quindi:

$$Cov((\mathbb{Z}^T\mathbb{Z})^{-1}\mathbb{Z}^T\vec{y}) = (\mathbb{Z}^T\mathbb{Z})^{-1}\mathbb{Z}^T \cdot Cov(\vec{y}) \cdot \left[(\mathbb{Z}^T\mathbb{Z})^{-1}\mathbb{Z}^T\right]^T = \sigma^2(\mathbb{Z}^T\mathbb{Z})^{-1}\mathbb{Z}^T\mathbb{Z}(\mathbb{Z}^T\mathbb{Z})^{-1} = \sigma^2(\mathbb{Z}^T\mathbb{Z})^{-1}$$

3)

$$\mathbb{E}\left[\widehat{\vec{\varepsilon}}\right] = \mathbb{E}\left[\vec{y} - \widehat{\vec{y}}\right] = \mathbb{E}\left[(\mathbb{1} - H)\vec{y}\right] = (\mathbb{1} - H)\mathbb{E}\left[\vec{y}\right] = (\mathbb{1} - H)\mathbb{Z}\vec{\beta} = 0$$

Uguale a zero perché $(\mathbb{1} - H)$ è il proiettore sul sottospazio ortogonale a $span(\mathbb{Z})$

4)

$$Cov(\widehat{\vec{\varepsilon}}) = (\mathbb{1} - H)Cov(\vec{y})(\mathbb{1} - H)^T = \sigma^2(\mathbb{1} - H)(\mathbb{1} - H)^T = \sigma^2(\mathbb{1} - H)^2 = \sigma^2(\mathbb{1} - H)$$

Perché $(\mathbb{1} - H)$ è un proiettore

5)

$$\mathbb{E}\left[\hat{\vec{\varepsilon}}^T\hat{\vec{\varepsilon}}\right] = \mathbb{E}\left[\sum_{i=1}^{n}\hat{\varepsilon}_i^2\right] = \mathbb{E}\left[tr\left(\hat{\vec{\varepsilon}}\hat{\vec{\varepsilon}}^T\right)\right]$$

Questo vale perché la traccia da la somma degli elementi sulla diagonale e se moltiplico una matrice per la trasposta ottengo tutte le componenti al quadrato

$$\mathbb{E}\left[tr\left(\hat{\vec{\varepsilon}}\hat{\vec{\varepsilon}}^T\right)\right] = \mathbb{E}\left[tr\Big((\mathbb{1}-H)\vec{y}[(\mathbb{1}-H)\vec{y}]^T\Big)\right]$$

Dato che $\vec{y} = \mathbb{Z}\vec{\beta} + \vec{\varepsilon}$ allora $(\mathbb{1}-H)\vec{y} = (\mathbb{1}-H)\vec{\varepsilon}$

$$\mathbb{E}\left[\hat{\vec{\varepsilon}}^T\hat{\vec{\varepsilon}}\right] = \mathbb{E}\left[tr\Big((\mathbb{1}-H)\vec{\varepsilon}\vec{\varepsilon}^T(\mathbb{1}-H)\Big)\right] = \mathbb{E}\left[tr\Big((\mathbb{1}-H)\vec{\varepsilon}\vec{\varepsilon}^T\Big)\right]$$

$$= tr\Big(\mathbb{E}\left[(\mathbb{1}-H)\vec{\varepsilon}\vec{\varepsilon}^T\right]\Big) = tr(\mathbb{1}-H)\,Cov(\vec{\varepsilon}) = \sigma^2(n-r-1)$$

Perché la traccia di un proiettore è uguale alla dimensione del sottospazio su cui proietta

Quindi $\dfrac{\sum\hat{\varepsilon}_i^2}{n-r-1} = S^2$ è stimatore non distorto per σ^2 □

Teorema: Gauss-Markov.

$$\vec{Y} = \mathbb{Z}\vec{\beta} + \vec{\varepsilon} \quad rango(\mathbb{Z}) = r+1$$

$\implies$ $\forall\vec{c}\in\mathbb{R}^{r+1}$ $\vec{c}^T\hat{\vec{\beta}}_L S$ è lo stimatore di $\vec{c}^T\vec{\beta}$ che ha la varianza minima tra tutti gli stimatori che:

1) Sono non distorti per $\vec{c}^T\vec{\beta}$

2) Sono della forma $a_1Y_1 + ... + a_nY_n$ (lineare $\vec{Y}$)

Ovvero $\vec{c}^T\hat{\vec{\beta}}_{LS}$ è BLUE, best linear unbiased estimator per $\vec{c}^T\vec{\beta}$

Dim.

Non distorto dimostrato nel teorema precedente

$$\vec{c}^T\hat{\vec{\beta}}_{LS} = \underbrace{\vec{c}^T}_{1\times(r+1)}\ \underbrace{(\mathbb{Z}^T\mathbb{Z})^{-1}}_{(r+1)\times(r+1)}\ \underbrace{\mathbb{Z}^T}_{(r+1)\times n}\ \vec{Y}$$

Quindi ottengo una moltiplicazione $(1\times n)$ per $(n\times 1)$ che quindi darà soluzione lineare in Y □

Vediamo un caso particolare interessante con $\vec{c}$ un vettore di zeri e un uno alla riga j $\quad\vec{c}^T\hat{\vec{\beta}}_{LS} = \hat{\vec{\beta}}_j$

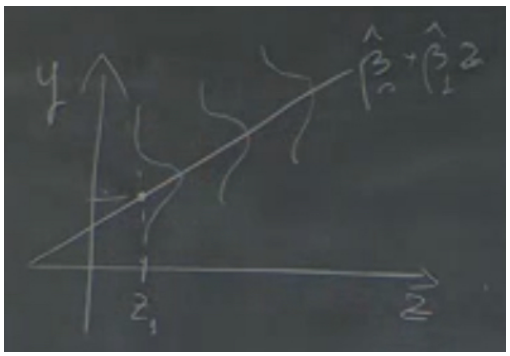
Quindi la componente j-esima che è uno stimatore unidimensionale ha varianza minima ed è non distorto

Per poter fare inferenza "puntuale" su $\vec{\beta}$ e $\vec{\sigma}$ (ovvero IC, test d'ipotesi ...) c'è bisogno di fare assunzione sulla legge di $\vec{\varepsilon}$

Quindi suppongo che $\vec{\varepsilon} \sim \mathcal{N}_n(\vec{0}, \sigma^2 \mathbb{1}) \iff \varepsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$ congiuntamente gaussiane

Avremo quindi che $\vec{Y} \sim \mathcal{N}_n(\mathbb{Z}\vec{\beta}, \sigma^2 \mathbb{1})$

Penserò y_1 come la realizzazione di una gaussiana di media $\beta_0 + \beta_1 z$ e varianza σ^2



Lezione 21 (05/05/23)

Teorema.

Assumiamo $rango(\mathbb{Z}) = r + 1$

I) $\hat{\vec{\beta}}_{LS}$ e $\hat{\sigma}^2 = \dfrac{\hat{\vec{\varepsilon}}^T \hat{\vec{\varepsilon}}}{n} = \dfrac{n-r-1}{n} S^2$ sono gli stimatori ML

II) $\hat{\vec{\beta}} \sim \mathcal{N}_{r+1}(\vec{\beta}, \sigma^2 (\mathbb{Z}^T \mathbb{Z})^{-1})$

III) $\hat{\vec{\varepsilon}} \sim \mathcal{N}_n(\vec{0}, \sigma^2 (\mathbb{1} - H))$

IV) $\hat{\vec{\beta}} \perp\!\!\!\perp \hat{\vec{\varepsilon}}$ indipendenti stocasticamente

V) $n\hat{\sigma}^2 = \hat{\vec{\varepsilon}}^T \hat{\vec{\varepsilon}} \sim \sigma^2 \cdot \chi^2(n-r-1)$

Dim.

I)

$$L = \prod_{i=1}^{n} f_{Y_i}(y_i) = \prod_{i=1}^{n} \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right)^n exp \left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} \left(y_i - \sum_j \beta_j z_{ij} \right)^2 \right\}$$

$$\varepsilon_i \sim \mathcal{N}(0, \sigma^2) \qquad Y_i \sim \mathcal{N}\left(\sum_j \beta_j z_{ij}, \sigma^2 \right)$$

Data la forma di L (funzione likelihood) massimizzare L equivale a massimizzare $\sum_{i=1}^{n}\left(y_i - \sum_{j=0}^{r}\beta_j z_{ij}\right)^2$

ma questo equivale a massimizzare i minimi quadrati $\implies \widehat{\vec{\beta}}_{ML} = \widehat{\vec{\beta}}_{LS}$

Dimostriamo i punti II)...IV) insieme:

$$\widehat{\vec{\beta}} = (\mathbb{Z}^T\mathbb{Z})^{-1}\mathbb{Z}^T\vec{Y} \quad \widehat{\vec{\varepsilon}} = (\mathbb{1} - H)\vec{Y}$$

$$\implies \begin{bmatrix} \widehat{\vec{\beta}} \\ \widehat{\vec{\varepsilon}} \end{bmatrix} = \begin{bmatrix} (\mathbb{Z}^T\mathbb{Z})^{-1}\mathbb{Z}^T \\ 1 - \mathbb{Z}(\mathbb{Z}^T\mathbb{Z})^{-1}\mathbb{Z}^T \end{bmatrix}$$

Quindi il vettore $\begin{bmatrix} \widehat{\vec{\beta}} \\ \widehat{\vec{\varepsilon}} \end{bmatrix}$ è un vettore gaussiano, essendo trasformazione lineare del vettore gaussiano $\vec{Y}$

$$\begin{bmatrix} \widehat{\vec{\beta}} \\ \widehat{\vec{\varepsilon}} \end{bmatrix} \sim \mathcal{N}_{n+r+1}\left(\begin{pmatrix} \vec{\beta} \\ \vec{0} \end{pmatrix}; \sigma^2 \begin{bmatrix} (\mathbb{Z}^T\mathbb{Z})^{-1} & (\mathbb{Z}^T\mathbb{Z})^{-1}\mathbb{Z}^T(1-H)^T = 0 \\ 0 & (\mathbb{1} - H) \end{bmatrix}\right)$$

$\implies \widehat{\vec{\beta}}$ e $\widehat{\vec{\varepsilon}}$ sono vettori scorrelati di un vettore gaussiano $\implies \widehat{\vec{\beta}} \perp\!\!\!\perp \widehat{\vec{\varepsilon}}$

IV) Ricordiamo che se $\vec{X} \sim \mathcal{N}(\vec{\mu}, \Sigma)$ dove Σ ha rango k, allora $(\vec{X} - \vec{\mu})^T\Sigma^{-1}(\vec{X} - \vec{\mu}) \sim \chi^2(k)$

$\widehat{\vec{\varepsilon}} \sim \mathcal{N}(\vec{0}, \sigma^2(\mathbb{1} - H)) \implies \widehat{\vec{\varepsilon}}^T \dfrac{(\mathbb{1} - H)^-}{\sigma^2}\widehat{\vec{\varepsilon}} \sim \chi^2(n - r - 1)$ ricordo A^- è la pseudo-inversa

Inoltre vale che $H\widehat{\vec{\varepsilon}} = 0$ e $\widehat{\vec{\varepsilon}} = (\mathbb{1} - H)\widehat{\vec{\varepsilon}}$

$$\widehat{\vec{\varepsilon}}^T \frac{(\mathbb{1} - H)^-(\mathbb{1} - H)}{\sigma^2}\widehat{\vec{\varepsilon}} = \frac{\widehat{\vec{\varepsilon}}^T\widehat{\vec{\varepsilon}}}{\sigma^2} \sim \chi^2(n - r - 1)$$

$$\implies \begin{cases} \widehat{\vec{\beta}} \sim \mathcal{N}(\vec{\beta}, \sigma^2(\mathbb{Z}^T\mathbb{Z})^{-1}) \\ \widehat{\vec{\varepsilon}}^T\widehat{\vec{\varepsilon}} \sim \sigma^2\chi^2(n - r - 1) \end{cases}$$

□

Corollario:

$$\begin{cases} \dfrac{1}{\sigma^2}(\widehat{\vec{\beta}} - \vec{\beta})^T(\mathbb{Z}^T\mathbb{Z})(\widehat{\vec{\beta}} - \vec{\beta}) \sim \chi^2(r+1) \\ \dfrac{n\widehat{\sigma}^2}{\sigma^2} \sim \chi^2(n - r - 1) \end{cases}$$ e sono tra loro indipendenti

La dimostrazione è conseguenza evidente del teorema e sono indipendenti perché dipendono rispettivamente solo da $\widehat{\vec{\beta}}$ e $\widehat{\vec{\varepsilon}}$ che sono indipendenti

Tutto ciò perché voglio costruire una regione di confidenza e quindi mi serve una quantità pivotale, adesso mi devo liberare di r dentro le χ^2

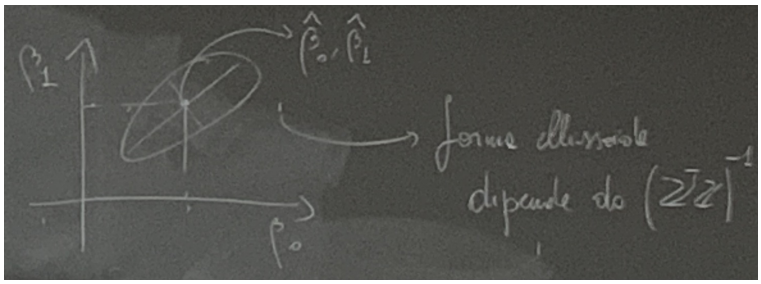
$$\frac{(\widehat{\vec{\beta}} - \vec{\beta})^T(\mathbb{Z}^T\mathbb{Z})(\widehat{\vec{\beta}\cdot} - \vec{\beta})\frac{1}{r+1}}{n\widehat{\sigma}^2 \cdot \frac{1}{n+r+1}} \sim F(r+1, n-r-1)$$

Ho ottenuto una quantità pivotale per $\vec{\beta}$ $\quad \dfrac{(\widehat{\vec{\beta}} - \vec{\beta})^T(\mathbb{Z}^T\mathbb{Z})(\widehat{\vec{\beta}\cdot} - \vec{\beta})\frac{1}{r+1}}{S^2} \sim F(r+1, n-r-1)$

Possiamo quindi calcolare le regioni di confidenza

$$\left\{\vec{\beta} \in \mathbb{R}^{n+1} \ (\widehat{\vec{\beta}} - \widehat{\beta})^T(\mathbb{Z}^T\mathbb{Z})(\widehat{\vec{\beta}} - \vec{\beta}) \leq (r+1)S^2F_{1-\alpha}(r+1, n-r-1)\right\}$$

É una regione di confidenza di livello $(1-\alpha)$ per $\vec{\beta}$ e ha forma di ellissoide



Sappiamo che $\widehat{\beta}_j \sim \mathcal{N}(\beta_j, \sigma^2(\mathbb{Z}^T\mathbb{Z})^{-1}_{jj})$

$$\implies \frac{\widehat{\vec{\beta}} - \vec{\beta}}{\sqrt{S^2(\mathbb{Z}^T\mathbb{Z})^{-1}_{jj}}} \sim t(n-r-1)$$

$IC_{1-\alpha}$ per β_j è $\left[\widehat{\beta}_j \pm \underbrace{\sqrt{S^2(\mathbb{Z}^T\mathbb{Z})^{-1}_{jj}}}_{se(\widehat{\beta_j})} t_{1-\frac{\alpha}{2}}(n-r-1)\right]$ dove $se(\widehat{\beta}_j)$ è lo standard error

Un altro metodo per trovare un intervallo di confidenza è usare il test d'ipotesi $\begin{cases} H_0 : \beta_j = 0 \\ H_1 : \beta_j \neq 0 \end{cases}$

$$\text{Sotto } H_0 \quad T_j = \frac{\widehat{\beta}_j}{\sqrt{S^2(\mathbb{Z}^T\mathbb{Z})^{-1}_{jj}}} \sim t(n-r-1)$$

$$RC = \left\{|T_j| > t_{1-\frac{\alpha}{2}}(n-r-1)\right\}$$

Quindi, posta t_j la realizzazione del test, posso calcolare il p-value del test $\mathbb{P}(T > |t_j|)$

Questo test d'ipotesi (con $\beta_j = 0$) equivale a fare la feature selection marginale, se ho evidenza che $\beta_j = 0$ allora posso trascurare la covariata j

Regressione su R

Vediamo cosa è importante guardare dal summary della regressione di R, cioè risultato del fit:

	Estimate	$se(\widehat{\beta}_j)$	$t-value$	$p-value$
β_0	Stima puntuale	$\sqrt{S^2(\mathbb{Z}^T\mathbb{Z})^{-1}_{jj}}$	$\frac{\text{stima puntuale}}{se(\widehat{\beta}_j)}$	...
β_1				
$\vdots$				

Il $t-value$ è il valore statistico del test $(\beta_j = 0 \; vs \; \beta_j \neq 0)$ sotto H_0

Quindi partendo da $\vec{Y} = \beta_0 + \beta_1\vec{Z}_1 + ... + \beta_r\vec{Z}_r + \vec{\varepsilon}$
Guarderò i p-value del summary e se quello corrispondente al test su β_j è alto, allora la covariata $\vec{Z}_j$ non è significativa e la posso togliere dal test

Posso fare anche i test analoghi sugli intervalli di confidenza, usando gli intervalli chiamati "one at a time" che sono intervalli marginali sulle singole covariate: $IC = \left[\widehat{\beta}_j \pm \sqrt{S^2(\mathbb{Z}^T\mathbb{Z})^{-1}_{jj}}\; t_{1-\frac{\alpha}{2}}(n-r-1)\right]$

- Un'altra cosa importante da guardare nel summary, oltre al p-value è il segno della stima puntuale, perché se ho una covariata continua e il β_j positivo, vuol dire che al crescre di quella covariata cresceranno i valori della risposta e questo ci servirà per spiegare la relazione tra la covariata e la variabile

- Altro fattore da guardare è l'ordine di grandezza, potrei avere un β_j molto significativo, ma se ha un ordine 10^{-3} questo ci dice l'impatto della covariata sulla risposta, per capire quant'è l'effetto dovremo vedere dove si muove la Z_j e dove si muove la Y

- Invece nel caso in cui le covariate hanno ordini di grandezza molto diversi, per evitare fastidi, si standardizzano le covariate

Lezione 22 (08/05/23)

Vediamo altri test sui parametri

$$\begin{cases} H_0 : C\vec{\beta} = 0 \\ H_1 : C\vec{\beta} \neq 0 \end{cases} \qquad C \text{ matrice } p \times (r+1)$$

Tra tutti questi test si usano test per vedere se ho delle covariate significative:

$$\begin{cases} H_0 : \beta_1 = \beta_2 = ... = 0 \\ H_1 : \exists j \text{ tc } \beta_j \neq 0 \end{cases}$$

Sfruttando test di questo tipo, posso costruire un test per confrontare il modello $\mathbb{Z}$ con un sotto modello $\mathbb{Z}_1$ in cui cancellop alcune covariate

$$\vec{Y} = \mathbb{Z}\vec{\beta} + \vec{\varepsilon} \qquad \vec{Y} = \mathbb{Z}_1\vec{\beta}_1 + \vec{\varepsilon}_1$$
$$SS_{res}(\mathbb{Z}) = \hat{\vec{\varepsilon}}^T\hat{\vec{\varepsilon}} = \sum_{i=1}^{n} \hat{\varepsilon}_i^2 \qquad SS_{res}(\mathbb{Z}_1) = \hat{\vec{\varepsilon}}_1^T\hat{\vec{\varepsilon}}_1 = \sum_{i=1}^{n} \hat{\varepsilon}_{i1}^2$$

Per confrontare i due modelli posso confrontare gli SS_{res}

So che vale sempre $SS_{res}(\mathbb{Z}) \leq SS_{res}(\mathbb{Z}_1)$ perché aggiungendo covariate, aumento la parte spiegata dal modello e quindi abbassare la parte residua, ovvero la parte che non riesco a spiegare

Se la differenza tra le due è grande, vuol dire che tra le due ho perso molto, se invece è piccola vuol dire che aggiungendo le covariate non cambia di molto la parte spiegata

$$\begin{cases} H_0 : \mathbb{Z} \text{ e } \mathbb{Z}_1 \text{ sono equivalenti} \\ H_0 : \mathbb{Z} \text{ e } \mathbb{Z}_1 \text{ non sono equivalenti} \end{cases}$$

"Rifiuto H_0 se $SS_{res}(\mathbb{Z}_1) - SS_{res}(\mathbb{Z})$ è grande"

Si può dimostrare che

$$\frac{SS_{res}(\mathbb{Z}_1) - SS_{res}(\mathbb{Z})}{S^2 \cdot p} \sim F(p, n-r-1)$$

Dove $S^2 = \dfrac{SS_{res}(\mathbb{Z})}{n-r-1}$ e p è la differenza tra le colonne $\mathbb{Z}$ e $\mathbb{Z}_1$

Rifiuto se $\dfrac{SS_{res}(\mathbb{Z}_1) - SS_{res}(\mathbb{Z})}{S^2 p} > F_{1-\alpha}(p, n-r-1)$

Se ottengo p-value basso significa sotto modello diverso, invece p-value alto significa sotto modello equivalente e quindi scelgo il sotto modello

Avere $\mathbb{Z}_1 = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}$ equivale al test $\begin{cases} H_0 : \beta_1 = ... = \beta_r = 0 \\ H_1 : \exists \beta_j = 0 \end{cases}$ questo è il test F che fa R in automatico

8.2 Anova

La tecnica Anova, ovvero analysis of varianze, è un modello lineare in cui le covariate sono solo categoriche, valuteremo questi gruppi per confrontare le medie

Esempio di confronto tra gruppi: supponiamo di avere 3 gruppi di costo appartamenti: "centro", "periferia" e "fuori città" e voglio valutare se il costo è lo "stesso" nei gruppi

Andrò a costruire una matrice, dove l'elemento $z_{ij} = \begin{cases} 1 \text{ se } i \in \text{ gruppo j} \\ 1 \text{ se } i \notin \text{ gruppo j} \end{cases}$

Permutando le righe della matrice di modo da avere prima le righe delle unità statistiche i nei primi gruppi, ottengo la matrice

$$\mathbb{Z} = \begin{bmatrix} 1 & 1 & & 0 & & 0 \\ \vdots & \vdots & \} n_1 & \vdots & \dots & \vdots \\ 1 & 1 & & 0 & & 0 \\ 1 & 0 & 1 & & & 0 \\ \vdots & \vdots & \vdots \} n_2 & & \dots & \vdots \\ 1 & 0 & 1 & & & 0 \\ \vdots & \vdots & \vdots & & & \vdots \\ 1 & 0 & 0 & & 1 & \\ \vdots & \vdots & \vdots & \dots & \vdots \} n_g & \\ 1 & 0 & 0 & & 1 & \end{bmatrix} \qquad \vec{Y} = \begin{bmatrix} X_{11} \\ \vdots \\ X_{n_1 1} \\ X_{12} \\ \vdots \\ X_{n_2 2} \\ \vdots \\ X_{1g} \\ \vdots \\ X_{n_g g} \end{bmatrix}$$

Quindi contando le righe n_j = numero di unità statistiche nel gruppo j e $\sum n_j = n$

Mentre nel vettore delle risposte $\vec{Y}$ gli X_{ij} sono la risposta dell'unità i del gruppo j, invece che scrivere $Y_1, ..., Y_n$

Posto, come al solito, $\vec{Y} = \mathbb{Z}\vec{\beta} + \vec{\varepsilon}$ allora $X_{ij} \sim \mathcal{N}(\mu_j, \sigma^2)$ con $\mu_j = \beta_0 + \beta_j$ perché $\mathbb{Z}\vec{\beta} = \begin{bmatrix} \beta_0 + \beta_1 \\ \vdots \\ \beta_0 + \beta_j \end{bmatrix}$

Quindi richiedere le medie uguali equivale a porre tutte le $\beta_j = 0$ e quindi che il sotto modello con solo l'intercetta sia equivalente al modello globale

Da questo si capisce che fittare un modello e quindi stimare la significatività dei β_j, in questo caso (covariate categoriche) equivale a fare un confronto tra le medie

Abbiamo un problema perché la matrice $\mathbb{Z}$ non ha rango massimo, infatti la prima colonna (intercetta) è la somma delle restanti. Quindi si usando delle matrici $\widetilde{\mathbb{Z}}$ a rango massimo che raccolgono le stesse informazioni, per esempio la seguente che è $n \times g$ invece che $n \times (g+1)$

$$\widetilde{\mathbb{Z}} = \begin{bmatrix} \mathbf{1} & \mathbf{1} & 0 & 0 \\ \vdots & 0 & \mathbf{1} & 0 \\ \vdots & \vdots & \vdots & \vdots \\ \vdots & 0 & 0 & \mathbf{1} \\ \mathbf{1} & \mathbf{-1} & \mathbf{-1} & \mathbf{-1} \end{bmatrix} \qquad \text{dove le colonne sono } g\text{, invece che } g-1$$

$$\mathbb{E}[X_{i1}] = \beta_0 + \beta_1 \qquad \mathbb{E}[X_{i2}] = \beta_0 + \beta_2 \quad ... \quad \mathbb{E}[X_{ig}] = \beta_0 - \beta_1 - ... - \beta_{g-1}$$

$$\mu_j = \beta_0 + \tau_j \qquad \sum \tau_j = 0$$

Quindi è una riparametrizzazione equivalente a prima, ma con il vincolo che la somma dei β sia nulla

Confrontiamo il modello $\widetilde{\mathbb{Z}}$ con il modello $\mathbb{Z}_1$

In $\mathbb{Z}_1$ ho solo l'intercetta e svolgendo il calcolo $\left(\mathbb{Z}_1(\mathbb{Z}_1^T\mathbb{Z}_1)^T\mathbb{Z}_1^T\right)$ ottengo che tutti i gruppi hanno la stessa media equivalente alla media su i e j di tutte le osservazioni

Quindi il valore fittato è: $\widehat{\vec{y}} = \begin{bmatrix} \overline{X}_{..} \\ \vdots \end{bmatrix}$

Dove $\overline{X}_{..} = \dfrac{1}{n}\sum_{ij} X_{ij}$ e supponendo $n_1 = ... = n_g = m$ allora $n = mg$ e $\overline{X}_{..} = \dfrac{1}{mg}\sum_{ij} X_{ij}$

Oss. Ricordo che le X_{ij} sono le Y_i, le avevo chiamate in modo diverso

Invece per $\widetilde{\mathbb{Z}}$, ovvero metodo $\mathbb{Z}$ che sono equivalenti, ottengo delle previsioni $\widehat{\widetilde{y}} = \begin{bmatrix} \left.\begin{matrix} \overline{X}_{\cdot 1} \\ \vdots \end{matrix}\right\} n_1 \\ \vdots \\ \left.\begin{matrix} \overline{X}_{\cdot g} \\ \vdots \end{matrix}\right\} n_g \end{bmatrix}$

Le previsioni sono uguali nei singoli gruppi, diverso da prima che erano uguali fra i gruppi

$\forall i \in$ gruppo j la previsione sarà $\overline{X}_{\cdot j} = \frac{1}{m} \sum_{i=1}^{m} X_{ij}$ dove abbiamo supposto per semplicità che $n_1 = ... = n_g = m$

Oss. Andremo a confrontare il residuo di quando prevedo la stessa media nei gruppi con il residuo di quando prevedo gruppo per gruppo, perché se questi due modelli sono equivalenti allora posso pensare che le medie dei gruppi siano le stesse, che è quello che voglio mostrare

Oss. Stiamo cercando di mostrare che le covariate, che sappiamo essere gaussiane, siano ugualmente distribuite e quindi vogliamo che abbiano stessa media

L'idea per confrontare i due modelli è usare il test sugli SS_{res}

$$\frac{\frac{SS_{res}(\mathbb{Z}_1) - SS_{res}(\widetilde{\mathbb{Z}})}{g-1}}{\frac{SS_{res}(\widetilde{\mathbb{Z}})}{mg-g}} = \frac{\frac{\sum_{ij}(X_{ij} - \overline{X}_{\cdot\cdot})^2 - \sum_{ij}(X_{ij} - \overline{X}_{\cdot j})^2}{g-1}}{\frac{\sum_{ij}(X_{ij} - \overline{X}_{\cdot j})^2}{mg-g}} \sim F(g-1, mg-g)$$

$$\sum_{j=1}^{g}\sum_{i=1}^{m}(X_{ij} - \overline{X}_{\cdot\cdot})^2 = \underbrace{\sum_{j=1}^{g}\sum_{i=1}^{m}(X_{ij} - X_{\cdot j})^2}_{SS_w} + \underbrace{m\sum_{j=1}^{g}(\overline{X}_{\cdot j} - \overline{X}_{\cdot\cdot})^2}_{SS_b}$$

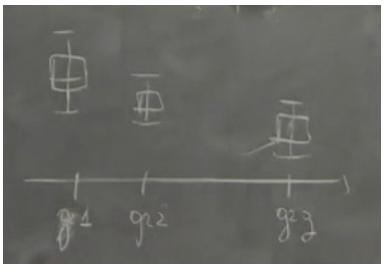
Dove abbiamo SS_{within} = variabilità nel gruppo e SS_{between} = variabilità tra i gruppi

Quindi il test diventa: $\frac{\frac{SS_b}{g-1}}{\frac{SS_w}{g(m-1)}} \sim F(g-1, g(m-1))$

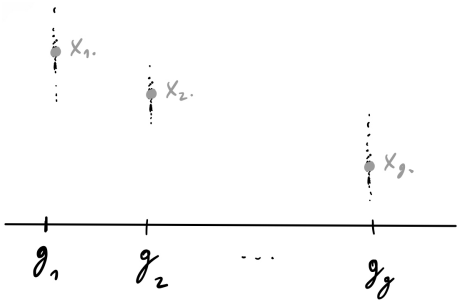
Rifiuto se la variabilità tra i gruppi è molto più grande della variabilità nei gruppi, perché questo da evidenza statistica che le medie nei gruppi sono diverse

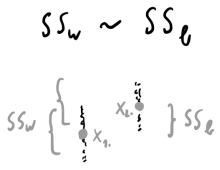
Vediamolo graficamente:

Senza questo metodo avrei per esempio potuto plottare i boxplot di ogni gruppo e graficamente vedere come erano distribuiti i gruppi



Invece questo metodo confronta le SS_w ovvero quante sono distribuite le nuvolette dei gruppi con la SS_b che è la variabilità tra le medie delle nuvolette.
Se questi valori hanno lo stesso ordine e quindi non ho nessun evidenza per dire che le medie sono diverse





Questo metodo funziona se $X_{ij} \sim \mathcal{N}(\mu_j, \sigma^2)$ e quindi oltre a verificare la gaussianità delle osservazioni dovrò verificare che abbiano varianza simile

Implementando Anova su R ricevo la Tabella Anova

Causa della variabilità	SS	$g.d.l.$	MS	F_0	$p-value$
Gruppi	SS_b	$g-1$	$\frac{SS_b}{g-1} = MS_b$	$\frac{MS_b}{MS_w}$	
Errore	SS_w	$g(m-1)$	$\frac{SS_w}{g(m-1)} = MS_w$		
Totale	SS_{tot}	$gm-1$			

Il p-value sarà la probabilità che la Fisher sia più grande della statistica che ottengo con i miei dati

Quindi con p-value piccolo rifiuto e quindi le medie tra i gruppi sono diverse, invece con p-value alto, accetto e le medie sono uguali

Oss. Ovviamente la differenza tra le medie non vuol dire che tutte le medie sono diverse

In conclusione, posta l'assunzione $X_{ij} \sim \mathcal{N}(\mu_j, \sigma^2)$, il test è $\begin{cases} H_0 : \mu_1 = ... = \mu_g \\ H_1 : \exists j \text{ tc } \mu_j \neq \mu_i \text{ per qualche i} \end{cases}$

Lezione 23 (15/05/23)

Principali finalità di questo tipo di modelli:

- Capire le relazioni tra le covariate e la risposta
- Fare previsione della risposta y in corrispondenza di valori $\vec{z}_i$, ovvero con una nuova osservazione o non appartenente al dataset

Spesso si **valida** un modello attraverso la CV, o cross validazione

A partire dal dataset $(y_i, \vec{z}_i)$, seleziono in modo randomico un sottoinsieme, che corrisponde a una certa percentuale del dataset, lo chiamo test set e lo metto da parte. A questo punto fitto il modello sul restante sottoinsieme che chiamerò training set.
Confronto le stime $\widehat{y}_i$ ottenute dal modello con solo il training per valutare la bontà della stima sul test set.

Posso ripetere la procedura un po' di volte e l'errore quadratico medio, ovvero la differenza tra il valore stimato e il valore vero sul test set, si chiama errore di cross validazione

Oss. Ovviamente non potrei confrontare confrontare il modello con dei dati che ho usato per costruire il modello

Supponiamo di avere una nuova osservazione $\vec{z}_0$ che non appartiene al dataset. Avrà covariate del tipo $\vec{z}_0 = (1, z_{0,1}, ..., z_{0r})^T$

Avrò due fonti di incertezza, quella data dagli $\widehat{\beta}_i$ e la variabilità σ^2
Posta Y_0 la risposta in corrispondenza di $\vec{z}_0$, allora $\mathbb{E}[Y_0] = \vec{z}_0^T \vec{\beta}$
Per stimare questa quantità uso il teorema di Gauss-Markov, per cui so che lo stimatore BLUE di $\vec{z}_0^T \vec{\beta}$ è $\vec{z}_0^T \widehat{\vec{\beta}} = \vec{z}_0^T (\mathbb{Z}^T \mathbb{Z})^{-1} \mathbb{Z}^T \vec{y}$ che mi fornisce una stima puntuale
Oltre a questa stima, vorrei costruire un intervallo di confidenza, per fare ciò devo capire la variabilità di

questo stimatore, ovvero che legge ha $\vec{z}_0 \widehat{\vec{\beta}}$

$\implies \vec{z}_0^T \sim \mathcal{N}(\vec{z}_0^T; \sigma^2 \vec{z}_0^T (\mathbb{Z}^T\mathbb{Z})^{-1} \vec{z}_0)$ perchè combinazione lineare di gaussiane

Voglio costruire una stima intervallare di questa, con la varianza incognita

Però so anche che $\dfrac{\widehat{\vec{\varepsilon}}^T \widehat{\vec{\varepsilon}}}{\sigma^2} \sim \chi^2(n-r-1)$ e che è indipendente da $\vec{z}_0^T \widehat{\vec{\beta}}$

$$\implies \frac{\vec{z}_0^T(\widehat{\vec{\beta}} - \vec{\beta})}{\sqrt{\vec{z}_0^T(\mathbb{Z}^T\mathbb{Z})^{-1}\vec{z}_0}\sqrt{S^2}} \sim t(n-r-1)$$

Posso quindi scrivere l'IC di livello $1-\alpha$ per $\mathbb{E}[y_0]$

$$IC = \left[\vec{z}_0^T \widehat{\vec{\beta}} \pm S\sqrt{\vec{z}_0^T(\mathbb{Z}^T\mathbb{Z})^{-1}\vec{z}_0} t_{1-\frac{\alpha}{2}}(n-r-1)\right] \quad \text{IC per la media di } Y_0$$

Nel caso $r = 1$, ovvero regressione lineare, e quindi $\vec{z}_0 = (1, z_0)^T$

Facendo il conto esplicito sulla parte di variabilità (larghezza) dell'IC, si ottiene

$$\sqrt{S^2\left[\frac{1}{n} + \frac{(z_0 - \overline{z})^2}{SS_z}\right]} \quad \text{dove } SS_Z = \sum (z_i - \overline{z})^2$$

Questo mi dice che la variabilità del mio modello dipende principalmente da quanto la mia osservazione $\vec{z}_0$ è lontana dal baricentro

Se invece voglio fare previsione sul valore puntuale di $\widehat{Y}_0$, questa previsione ha incertezza che ho nello stimare la media e poi un'incertezza legata a σ^2, ovvero la variabilità della gaussiana

Dato che $Y_0 = \vec{z}_0^T \vec{\beta} + \varepsilon_0$

Svolgendo una procedura analoga a prima e ricordandosi che ε_0 è indipendente dai β, si ottiene

$$\frac{Y_0 - \vec{z}_0^T \widehat{\vec{\beta}}}{\sqrt{S^2(1 + \vec{z}_0^T(\mathbb{Z}^T\mathbb{Z})^{-1}\vec{z}_0)}} \sim t(n-r-1)$$

Questa è una t-student perché è il quoziente di una gaussiana standardizzata e la radice di una Chi-quadro divisa per i suoi gradi di libertà, però standardizzando questa gaussiana devo tenere conto della variabilità dei $z\beta$ e di ε e per questo al denominatore devo aggiungere un S^2

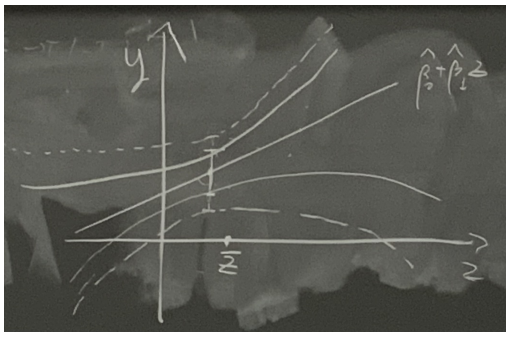
Quindi posso dire che:

$$\mathbb{P}\left(Y_0 \in \left[\vec{z}_0^T \widehat{\vec{\beta}} \pm \sqrt{S^2(1 + \vec{z}_0^T(\mathbb{Z}^T\mathbb{Z})^{-1}\vec{z}_0)}\; t_{1-\frac{\alpha}{2}}(n-r-1)\right]\right) = 1-\alpha \quad \text{intervallo di previsione di } Y_0$$

Ovvero la probabilità che una t-student stia tra i suoi quantili

L'intervallo per Y_0 sarà più ambio di quello della media di Y_0 perché tiene conto di ε_0

Nel caso di regressione lineare semplice possiamo visualizzare questa cosa:



Nella figura vediamo il modello fittato $(\widehat{\beta}_0 + \widehat{\beta}_1 z)$
Inoltre per ogni valore possiamo calcolare un intervallo di confidenza per la media, centrato sulla retta e l'intervallo di previsione centrato nello stesso punto, ma con larghezza maggiore
Posso sviluppare gli inviluppi di questi intervalli di confidenza. Per la media (—) per la previsione (- - -)
Questi inviluppi si allargano perché tanto più mi allontano dalla media $\overline{Z}$, tanto più l'ampiezza aumenta

Questo è sensato perché se sono vicino alla media vuol dire che sono vicino ai punti che ho usato poer fittare il modello. Questo fenomeno si chiama problemi di estrapolazione

Vediamo un esempio di problema di estrapolazione, che evidenzia l'errore al di fuori della "finestra" in cui osservo i dati

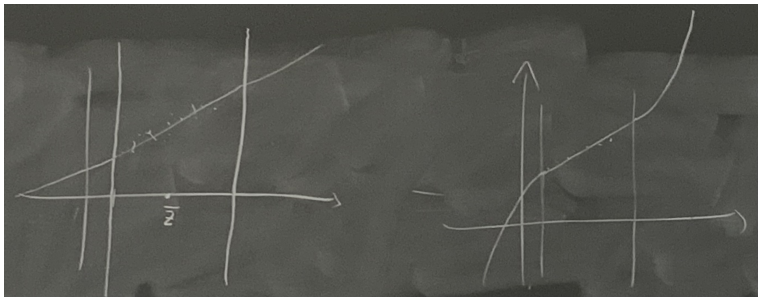


Figure 2: regressione valori reali

Cerchiamo di capire cosa succede quando ho un modello che comprende covariate sia continue che categoriche

Esempio: Y=stipendio

Z_1 = anni di servizio (continua)

Z_2 = sesso (binaria)

$$\vec{Y} = \beta_0 + \beta_1 Z_1 + \beta_2 Z_2 + \varepsilon$$

$\forall i$ tc $z_{2i} = 0$ (maschi) $\quad y = \widehat{\beta_0} + \widehat{\beta}_1 Z_1$

$\forall i$ tc $z_{2i} = 1$ (femmine) $\quad y = \widehat{\beta_0} + \widehat{\beta}_1 Z_1 + \widehat{\beta}_2$

Però è un po' limitante questo modello perché è possibile che ci sia una relazione tra Z_1 e Z_2 che quindi implica una diversa pendenza

Per esempio in questo caso posso avere un aumento di guadagno minore per le donne con l'aumento degli anni di servizio

Per poter valutare questa cosa, sempre con un modello lineare, introduco l'interazione, ovvero una nuova covariata

Oss. La linearità è nei β non negli Z

$$Y = \beta_0 + \beta_1 Z_1 + \beta_2 Z_2 + \beta_3 Z_1 Z_2 + \varepsilon$$

$\forall i$ tc $z_{2i} = 0$ (maschi) $\quad y = \widehat{\beta_0} + \widehat{\beta}_1 Z_1$

$\forall i$ tc $z_{2i} = 1$ (femmine) $\quad y = \widehat{\beta_0} + (\widehat{\beta}_1 + \widehat{\beta}_3) Z_1 + \widehat{\beta}_2$

Posso fare un test $\begin{cases} H_0 : \beta_3 = 0 \\ H_1 : \beta_3 \neq 0 \end{cases}$ che se ha un p-value basso, allora c'è interazione e quindi le due rette possono avere pendenza diversa, in particolare la stima di β_3 ci stima la differenza nello slope delle due curve

Oss. β_3 è la differenza tra i coefficienti angolari al passare da la categoria $z_2 = 0$ a $z_2 = 1$, vale lo stesso con più variabili categoriche, solamente avrei più coefficienti di interazione e il coefficiente angolare sarebbe Δy all'aumentare di 1 della variabile continua

Vale la stessa cosa nel caso di variabili continue:

$$Z_1, Z_2 \text{ continue } \quad \vec{Y} = \beta_0 + \beta_1 \vec{Z}_1 + \beta_2 \vec{Z}_2 + \beta_3 \vec{Z}_1 \vec{Z}_2 + \vec{\varepsilon}$$

$$Y(Z_1) = \beta_0 + \beta_1 Z_1 + \beta_2 Z_2 + \beta_3 Z_1 Z_2 + \varepsilon$$

$$Y(Z_1 + \delta) = \beta_0 + \beta_1 (Z_1 + \delta) + \beta_2 Z_2 + \beta_3 (Z_1 + \delta) Z_2 + \varepsilon$$

$$\Delta Y = Y(Z_1 + \delta) - Y(Z_1) = \beta_1 \delta + \beta_3 Z_2 \delta$$

Quindi se c'è interazione, al variare dell'osservazione per Z_1, la Y varia anche per causa di Z_2

Punti di leva:

Oss. Visti a laboratorio, quindi li vediamo velocemente

Nel nostro modello

$$\text{Var}\left[\vec{\varepsilon}\right] = \sigma^2 \mathbb{1} \quad \text{Var}\left[\varepsilon_i\right] = \sigma^2$$

Fittando il modello, ottengo

$$\text{Var}\left[\widehat{\vec{\varepsilon}}\right] = \sigma^2 (\mathbb{1} - H) \quad \text{Var}\left[\widehat{\varepsilon}_i\right] = \sigma^2 (1 - h_{ii})$$

h_{ii} è l'elemento i-esimo della diagonale di H, si chiama leverage del dato i

Per esempio se $h_{ii} = 1$ allora $\text{Var}\left[\widehat{\varepsilon}_i\right] = 0$, ovvero sto interpolando il dato

Quindi tanto più la leverage tende a 1, tanto più il modello è forzato a passare per quel dato, ovvero fa effetto leva

Vediamolo in modo più generale:

$$\widehat{\vec{y}} = \mathbb{Z} \widehat{\vec{\beta}} = H \vec{y}$$

$$\widehat{y}_i = \sum_{j=1}^{n} h_{ij} y_j = h_{ii} y_i + \sum_{\substack{j=1 \\ j \neq 1}}^{n} h_{ij} y_j$$

Però $H \vec{1} = \vec{1}$, perché c'è l'intercetta in H, ma $H \vec{1} = \sum_{j=1}^{n} h_{ij} = 1$ e $\sum_i h_{ii} = tr(H) = r + 1$

Si fa quindi il Plot dei Leverages

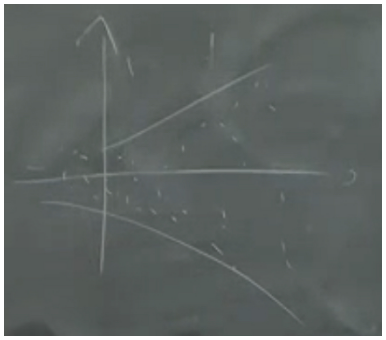
Però non voglio avere interpolazione, ovvero non voglio avere dei punti con levarages più alti di altri, la mia situazione ideale sarebbe avere tutti i punti con lo stesso effetto leva, ovvero $h_{ii} = \dfrac{r+1}{n}$

Quindi si mette una soglia a $2\left(\frac{r+1}{n}\right)$ e tutte le osservazioni con $h_{ii} > 2\left(\frac{r+1}{n}\right)$ verranno chiamati punti leva e ci devo prestare attenzione, tendenzialmente li escluderò dal modello

Grafici di diagnostica:

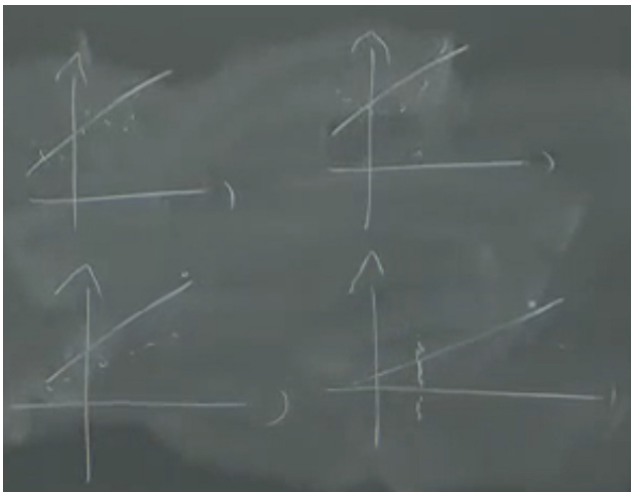
Una buona regola per vedere se il modello sta andando bene è fare questi grafici di diagnostica che nel pratico sono gli scatterplot degli $(\widehat{y}_i, \widehat{\varepsilon}_i)$ o $\widehat{y}_i, \frac{\widehat{\varepsilon}_i}{\sqrt{1-h_{ii}}}$

Questa nuvola di punti deve essere il più possibile casuale, centrata sullo zero, Non voglio vedere un plot a imputo che si allarga in avanti o all'indietro, perché questo tipo di grafico viola la omoschedasticità, ovvero la varianza cambia al variare della y_i



Esempio del quartetto di Hanscobe:

Si generano delle regressioni con stessi R^2 e $\widehat{\beta}_0, \widehat{\beta}_1$, ma sono generate da dataset molto diversi



L'idea è che guardare solo R^2 e i $\widehat{\beta}$ non ci dice niente su quanto il modello vada bene, dovrò sempre fare un grafico di diagnostica dei residui

Normalità dei residui: Per valutare la normalità si usa uno shapiro test o qqplot
Che mi confrontano i quantili di una gaussiana, rispetto ai quantili del mio test

Lezione 24 (16/05/23)

Procediamo con **Indici di goodness of fitness:**

Avevamo visto che:
R^2 mi restituisce la quota parte del dataset spiegata dal modello $R^2 = \frac{SS_{reg}}{SS_{tot}} = 1 - \frac{SS_{res}}{SS_{tot}}$

Possiamo anche dover confrontare modelli con un numero di covariate diverso
Supponiamo di avere un modello M_h con h covariate e un sovra modello M_k con $k > h$
Un confronto di R^2 non è "fair", perché ovviamente si alzerà. Per mitigare l'aumento di R^2 dovuto all'aumento delle covariate, si introduce R^2_{adj} (adjusted) tale che $1 - R^2_{adj} = \frac{\frac{SS_{res}}{n-r-1}}{\frac{SS_{tot}}{n-1}}$ ovvero $R^2_{adj} = 1 - (1 - R^2)\frac{n-1}{n-r-1}$

Metodi per selezionare le covariate:

Supponiamo di avere r covariate, il metodo ottimale per decidere quali covariate togliere sarebbe confrontare tutti i possibili sottomodelli
Se voglio un sottomodello con:

1 covariata, avrò r modelli possibili

h covariate, avrò $\binom{r}{h}$ possibili

r covariate avrò 1 modello

Però appena h aumenta, confrontare tutti i modelli diventa computazionalmente poibitivo

Si usano i **Metodi steopwise:**

a) foward selection, che aggiunge covariate

b) backward selection, che parte con tutte le covariate e ne toglie una alla volta

c) both che aggiunge e toglie

Per decidere quando fermarmi, ad ogni iterazione, si confronta il modello $M_k(i)$, con i indice di iterazione e k numero di covarite, con $M_{k'}(i+1)$ e si fa il test F

Se $k' > k$ allora si valuta $\frac{\frac{SS_{res}(M_k(i)) - SS_{res}(M_{k'}(i+1))}{k'-k}}{\frac{SS_{res}(M_k(i))}{n-k-1}}$ e se il p-value è alto allora i modelli sono equivalenti

Collinearità tra le covariate:

Se la collinearità è alta, allora la varianza $\sigma^2(\mathbb{Z}^T\mathbb{Z})^{-1}$ "esplode"

Per controllare la collinearità, c'è un indice che si chiama **VIF** = variance inflation factor che per ogni covariata j vale $VIF_j = \frac{1}{1-R_j^2}$ dove R_j^2 è il coefficiente di determinazione di un modello in cui si pone la covariata $\vec{Z}_j$ come risposta e le altre covariate come predittori, nel senso che mi dimentico di Y e considero la covariata come risposta. A questo punto tanto più R_j^2 è alto, tanto più $\vec{Z}_j$ sarà combinazione lineare delle altre covariate e quindi VIF è alto e quindi σ^2 rischia di esplodere

Vediamo come si risolvono alcuni problemi e come si identificano gli outlier

Problemi di eteroschedasticità:

Che si verificano quando non ho omoschedasticità, ovvero che l'assunzione $Cov(\vec{\varepsilon}) = \sigma^2\mathbb{1}$ non è verificata

E quindi $Cov(\vec{\varepsilon}) = \sigma^2\Sigma$ ovvero è una matrice più complicata

In questo caso esiste $C : CC^T = C^TC = \Sigma^{-1}$ con $C = \sqrt{\sigma^{-1}}$, detta radice di Σ

Se riscriviamo il modello come $C\vec{Y} = C\mathbb{Z}\vec{\beta} + C\vec{\varepsilon}$ allora si ottiene Var $[C\vec{\varepsilon}] = \sigma^2\mathbb{1}$
In questo caso avaremo $\hat{\vec{\beta}} = (\mathbb{Z}^T\Sigma^{-1}\mathbb{Z})^{-1}(\mathbb{Z}^T\Sigma^{-1}\vec{Y})$

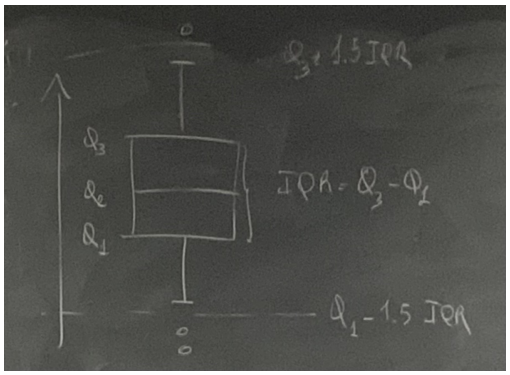
Esempio:

Caso in cui la matrice Σ è simile all'identità, ma diversa $\Sigma = \sigma^2 \begin{pmatrix} m_1 & & \\ & \ddots & \\ & & m_n \end{pmatrix}$

I dati vengono trasformati con $\sqrt{\Sigma^{-1}}$ che è la diagonale con termini $1\frac{1}{\sqrt{\sigma^2 m_i}}$
Un esempio in cui ho la matrice di questo tipo, ho i dati che provengono da gruppi e la variabilità scala con la numerosità del gruppo
Allora stiamo trasformando le Y tenendo conto di queste radici. Per qeusto motivo si chiama regressione pesata

Outliers:

Costruisco il boxplot:
Creo la "scatola" centrata in Q_2, con estremi Q_1 e Q_3, pongo $IQR = Q_3 - Q_1$ come l'ampiezza della scatola
Poi le "righe" arrivano fino a $Q_1 - 1.5 * IQR$ e $Q_3 + 1.5 * IQR$
Tutti i dati che sono al di fuori di questi sono detti outliers e li segno con dei pallini



Costruita in questo modo, la probabilità di avere outliers per una gaussiana è 0.007

Non devo buttare via gli outliers, la presenza di outliers va indagata perché può avere diversi motivi

Punti influenti:

A partire dal modello $\vec{Y} + \mathbb{Z}\vec{\beta} + \vec{\varepsilon}$

Chiamiamo $\mathbb{Z}_{(i)}$ la matrice disegno a cui tolgo la riga i, è matrice $(n-1) \times (r+1)$

E $\vec{Y}_{(i)}$ il vettore delle risposte a cui ho tolto y_i

A questo punto fittiamo il modello $\vec{Y}_{(i)} + \mathbb{Z}_{(i)}\vec{\beta} + \vec{\varepsilon}_{(i)}$

Ottego $\hat{\vec{\beta}}_i = (\mathbb{Z}_{(i)}^T \mathbb{Z}_{(i)})^{-1} \mathbb{Z}_{(i)}^T \vec{Y}_{(i)}$

Dovrò avlutare la discrepanza tra $\hat{\vec{\beta}}$ e $\hat{\vec{\beta}}_{(i)}$ se sono distanti, allora il sarà un dato influente

Ovvero un dato è influente se togliendolo, mi cambiano le stime

Però questa è la "distanza" tra due realizzazioni di vettorre gaussiani, dovrò valutare la distanza tenendo anche conto della variabilità

Si usa:

Definizione.

La **distanza di Cook** $= D_i$ tra $\hat{\vec{\beta}}$ e $\hat{\vec{\beta}}_{(i)}$ è

$$D_i = \frac{(\hat{\vec{\beta}}_{(i)} - \hat{\vec{\beta}})^T \mathbb{Z}^T \mathbb{Z} (\hat{\vec{\beta}}_{(i)} - \hat{\vec{\beta}})}{S^2(r+1)}$$

Un D_i alto vuol dire che i è un punto molto influente

Questo valore è uguale se valuto la differeza tra $\hat{\vec{Y}} = \mathbb{Z}\hat{\vec{\beta}}$ e $\hat{\vec{Y}}_{(i)} = \mathbb{Z}\hat{\vec{\beta}}_{(i)}$, ovvero $D_i = \dfrac{(\hat{\vec{Y}}_{(i)} - \hat{\vec{Y}})^T (\hat{\vec{Y}}_{(i)} - \hat{\vec{Y}})}{S^2(r+1)}$

Si può anche dimostrare che $D_i = \dfrac{1}{r+1} \cdot \hat{\varepsilon}_i^* \left(\dfrac{h_{ii}}{1-h_{ii}} \right)$ dove $\hat{\varepsilon}_i^* = \dfrac{\hat{\varepsilon}_i}{S\sqrt{1-h_{ii}}}$ è il residuo studentizzato

E questo ci dice che distanze di Cook alte e leverages alte sono paragonabili

Caso in cui i residui non sono gaussiani:

Si trasformano le y_i per tirarli a una normale

Una trasformazione che spesso porta i dati a normalità e il logaritmo, nel senso che i dati sono distribuiti come una log normale. Questa trasformazione rientra in una categoria più grossa:

Trasformazioni Box-Cox, che sono trasformazioni polinomiali o logaritmniche:

$$y^{new} = \frac{(y^\lambda - 1)}{\lambda} \qquad y^{(new)} = log\, y$$

Calcolo λ^{opt} tale per cui la verosimiglianza gaussiana di y^{new} è massima

Nel pratico si passano i possibili lambda e con ognuno di questi si cerca quello più gaussiano

Se $\lambda^{opt} = 0$ allora uso il logaritmo $y^{(new)}$, altrimenti uso la trsformazione polinomiale y^{new}

Tipicamente r restituisce un profilo di verosimiglianza in funzione di λ

E l'ottimo è la migliore trasformazione tra quelle di questo tipo

Oss. Se ottengo, per esempio, $\lambda^{opt} = 0.372$ non userò questo λ perché è un valore che non sono in grado di motivare, prenderò $\sqrt[3]{y}$ e già questo è poco interpretabile

AIC, indice di Akaike:

Serve in generale a confrontare due modelli ed è definito $AIC = -2ln(L) + 2k$ con $k =$ numero di parametri del modello

Confrontando due modelli preferirò un AIC piccolo

9 Modelli lineari generalizzati (GML)

Lezione 25 (22/05/23)

Si vogliono estendere i modelli lineari a casi in cui la variabile dipendente Y non è necessariamente gaussiana, ma $Y \in EF$, famiglia esponenziale
Quindi studieremo $Y \sim$ Bernoulli, $Y \sim$ Poisson, $Y \sim$ multinomiale ...

Si vuole quindi modellare $\mathbb{E}[Y]$, o una sua funzione, in base alle covariate

Oss. Avremo un set di variabili per spiegare una funzione della media della risposta
Come al solito gli obbiettivi sono capire la relazione tra Y e Z oppure fare previsione

Sarà necessario specificare 3 componenti:
1) Random component, ovvero Y risposta dipendente aleatoria (specificarne quindi la legge)
2) Componente sistematica, ovvero $\mathbb{Z}$ matrice dei predittori
3) Link function $g(\cdot)$ che specifica quale funzione della media voglio modellare con $\mathbb{Z}$

Per la famiglia esponenziale possiamo scrivere $f(y_i, \theta_i) = a(\theta_i)b(y_i)\, exp[y_i Q(\theta_i)]$
Dove $Q(\theta_i)$ è detta parametro naturale
Oss. Spesso sceglieremo la link function $g \equiv Q$

Vediamo i casi più comuni:

$$Y_i \sim Be(p_i) \qquad p_i = \mathbb{E}[Y_i]$$

$$f(y_i, p_i) = p_i^{y_i}(1-p_i)^{1-y_i}\mathbb{1}_{\{0,1\}}(y_i) = (1-p_i)\mathbb{1}_{\{0,1\}}(y_i)\, exp\left\{y_i\, log\left(\frac{p_i}{1-p_i}\right)\right\}$$

Parametro naturale per $y_i \sim Be(p_i)$ è $\quad log\left(\frac{p_i}{1-p_i}\right) = logit(p_i)$
Oss. Quindi andremo a studiare il $logit(\mathbb{E}[Y_i])$, inoltre $logit(\cdot) : [0,1] \to \mathbb{R}$. Questo ci risolve problemi di non invertibilità. La link serve invertibile per poter stimare i β a partire da stime di μ

$$Y_i \sim P(\lambda_i) \qquad f(y_i, \lambda_i) = \frac{e^{-\lambda}\lambda_i^{y_i}}{y_i!}\mathbb{1}_{\mathbb{N}}(y_i) = \frac{e^{-\lambda}\mathbb{1}_{\mathbb{N}}(y_i)}{y_i!}\, exp\{y_i\, log(\lambda_i)\}$$

$$\implies \text{ parametro naturale è } log(\lambda_i)$$

Anche qui avevamo λ_i solo positivo, però $log(\cdot) : [0, +\infty] \to \mathbb{R}$

La componente sistematica, n generale, sarà $\mathbb{Z}\vec{\beta} = (\eta_1, ..., \eta_n)^T \quad \mathbb{Z} = \begin{bmatrix} 1 & Z_{11} & \dots & Z_{1r} \\ \vdots & \vdots & \dots & \vdots \\ 1 & Z_{n1} & \dots & Z_{nr} \end{bmatrix}$

Dove $\eta_i = \sum_j \beta_j Z_{ij} = \beta_0 + \beta_{i1} + ... + \beta_r Z_{ir}$ predittore lineare

$\vec{\beta}$ è incognito e vorremo trovargli degli stimatori

Si vuole trovare link function tale che $g(\mu_i) = \eta_i = \beta_0 + \beta_{i1} + ... + \beta_r Z_{ir}$ dove $\mu_i = \mathbb{E}[Y_i]$

Tendenzialmente si sceglierà $g(\mu_i)$= parametro naturale, che viene chiamata link function canonica

Vediamo il caso della Bernoulli $Y_i \sim Be(p_i)$

Link function canonica è $logit(p_i)$

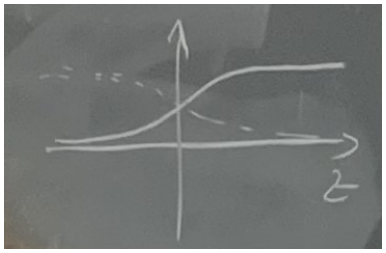
$$logit\left(\frac{p_i}{1-p_i}\right) = \eta_i = \beta_0 + \beta_{i1} + ... + \beta_r Z_{ir}$$

In questo caso $p_i \in [0, 1]$ mentre $logit(p_i) \in \mathbb{R}$ quindi si può invertire

$$logit(p_i) \iff log\left(\frac{p_i}{1-p_i}\right) = \eta_i \iff \frac{p_i}{1-p_i} = exp(\eta_i) \iff p_i = (1-p_i)\, exp(\eta_i) \iff p_i(1+exp(\eta_i)) = exp(\eta_i)$$

$$\implies p_i = \frac{exp[\beta_0 + \beta_1 Z_{i1} + ... + \beta_r Z_{ir}]}{1 + exp[\beta_0 + \beta_1 Z_{i1} + ... + \beta_r Z_{ir}]}$$

Nel caso in cui $r = 1$ avremo $p_i = \frac{exp[\beta_0 + \beta_1 Z_{i1}]}{1 + exp[\beta_0 + \beta_1 Z_{i1}]}$ che in base al segno di β_1 avrà andamento del tipo:



Il segno di β_1 dice come cambia la probabilità p_1 al variare della covariata

Oss. Se β_1 è positiva allora aumentando la covariata allora aumenterà la probabilità che la risposta sia 1

Vediamo il caso della Poisson $Y_i \sim P(\lambda_i)$

Avremo la link function canonica $g = log(\lambda_i)$

$$log(\lambda_i) = \eta_i \quad \lambda_i = exp\{\eta_i\} = \exp\{\beta_0 + \beta_1 Z_{i1} + ... + \beta_r Z_{ir}\}$$

9.1 Devianza

Oss. La devianza è l'indice più usato per confrontare modelli

Dati un modello lineare generalizzato $(y_1...y_n)^T$ e la log-likelihood $l(\vec{\mu}, \vec{y})$,pensata come funzione di $\vec{\mu}$
Supponiamo di avere i $\hat{\vec{\beta}}_{ML}$, allora $l(\hat{\vec{\mu}}, \vec{y})$ = log-likelihood del modello in cui uso gli stimatori $\hat{\vec{\mu}}$

$$\mu = g^{-1}(\eta_i) \implies \widehat{\mu} = g^{-1}(\widehat{\eta}_i) = g^{-1}(\widehat{\beta}_0 + \widehat{\beta}_1 Z_{i1} + ... + \widehat{\beta}_r Z_{ir})$$

Tra tutti i possibili modelli, ovvero tutti i possibili modi in cui mettere dentro $\hat{\vec{\mu}}$, la log-likelihood è massimizzata dal modello saturato ovvero $l(\vec{y}, \vec{y})$, dove come stima dei parametri ho posto le osservazioni

Tra tutti i possibili modelli, ovvero tutti i possibili modi in cui mettere dentro $\hat{\vec{\mu}}$, la log-likelihood è massimizzata dal modello saturato ovvero $l(\vec{y}, \vec{y})$, dove come stima dei parametri ho posto le osservazioni

Oss. La devianza sarà la differenza tra la log-likelihood massima e quella del modello

Definizione.
Devianza $= -2\big[l(\hat{\vec{\mu}}, \vec{y}) - l(\vec{y}, \vec{y})\big] = -log\big[$ rapporto tra likelihood modello fittato e modello saturato $\big]$

9.2 Exponential dispertion family

Vogliamo trovare un metodo per stimare i parametri incogniti $\vec{\beta}$

La famiglia EDF è del tipo $f(y_i, \theta_i, \phi) = exp\left[\frac{(y_i\theta_i - b(\theta_i))}{a(\phi)} + c(y_i, \phi)\right]$

Scrivendo in questo modo avremo che ϕ è il parametro di dispersione e θ_i è il parametro naturale
In questa famiglia abbiamo:

$$L = \prod_{i=1}^{n} f(y_i, \theta_i, \phi) \qquad log\, L = l = \sum_{i=1}^{n} l_i = \sum_{i=1}^{n} log\, f(y_i, \theta_i, \phi) \qquad l_i = \frac{[y_i\theta_i - b(\theta_i)]}{a(\phi)} + c(y_i, \phi)$$

$$\frac{\partial l_i}{\partial \theta_i} = \frac{y_i - b'(\theta_i)}{a(\phi)} \qquad \frac{\partial^2 l_i}{\partial^2 \theta_i} = -\frac{b''(\theta_i)}{a(\phi)}$$

$$\mathbb{E}\left[\frac{\partial l_i}{\partial \theta_i}\right] = 0 \iff \mathbb{E}\left[Y_i\right] = b'(\theta_i)$$

$$\mathbb{E}\left[-\frac{\partial^2 l_i}{\partial^2 \theta_i}\right] = \mathbb{E}\left[\left(\frac{\partial l_i}{\partial \theta_i}\right)^2\right] \implies \frac{b''(\theta_i)}{a(\phi)} = \mathbb{E}\left[-\frac{\partial^2 l_i}{\partial^2 \theta_i}\right] = \mathbb{E}\left[\frac{Y_i - b'(\theta_i)}{a(\phi)}\right] = \frac{1}{a(\phi)^2}\mathbb{E}\left[(Y_i - b'(\theta_i))^2\right] = \frac{\mathrm{Var}\left[Y_i\right]}{a(\phi)^2}$$

$$\text{In conclusione abbiamo: } \begin{cases} \mathbb{E}[Y_i] = b'(\theta_i) \\ \text{Var}[Y_i] = b''(\theta_i)a(\phi) \end{cases}$$

Applichiamo questo risultato nel caso della Poisson e della Bernoulli

$$Y_i \sim P(\mu_i) \quad f(y_i, \theta_i) = \frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!} = exp\{y_i\, log(\mu_i) - \mu_i - log(y_i!)\}$$
$$\implies \theta_i = log(\mu_i) \quad b(\theta_i) = exp(\theta_i)$$
$$\implies EY_i = b'(\theta_i) = exp(\theta_i) = \mu_i \quad \text{Var}[Y_i] = b''(\theta_i) = exp(\theta_i) = \mu_i$$

Oss. Quindi le Poisson hanno media e varianza uguali perché il parametro $b(\cdot)$ è l'esponenziale

$$Y_i \sim Be(p_i) \quad f(y_i, \theta_i) = p_i^{y_i}(1-p_i)^{1-y_i} = exp\left\{y_i\, log\left(\frac{p_i}{1-p_i}\right) + log(1-p_i)\right\} = exp\{y_i\theta_i - b(\theta_i)\}$$
$$\implies \theta_i = log\left(\frac{p_i}{1-p_i}\right) \quad log(1-p_i) = -log(1+exp(\theta_i)) = -log\left(1+\frac{p_i}{1-p_i}\right) \implies b(\theta_i) = log(1+exp(\theta_i))$$
$$\implies EY_i = b'(\theta_i) = \frac{exp(\theta_i)}{1+exp(\theta_i)} = p_i \quad \text{Var}[Y_i] = b''(\theta_i) = \frac{exp(\theta_i)}{(1+exp(\theta_i))^2} = \frac{exp(\theta_i)}{1+exp(\theta_i)} \cdot \frac{1}{1+exp(\theta_i)} = p_i(1-p_i)$$

Oss. tutto questo per ricavare cose che sapevamo già, ma almeno le abbiamo legate alla componente b della famiglia

Lezione 26 (23/05/23)

9.3 Equazioni di verosimiglianza

Per trovare gli stimatori usiamo le equazioni per trovare i massimi della verosimiglianza

$$\frac{\partial l}{\partial \beta_j} = \sum_{i=1}^{n} \frac{\partial l_i}{\partial \beta_j} = 0 \quad \forall j \qquad \text{con } l_i = \frac{[y_i\theta_i - b(\theta_i)]}{a(\phi)} + c(y_i, \phi)$$
$$\frac{\partial l_i}{\partial \beta_j} = \frac{\partial l_i}{\partial \theta_i} \cdot \frac{\partial \theta_i}{\partial \mu_i} \cdot \frac{\partial \mu_i}{\partial \eta_i} \cdot \frac{\partial \eta_i}{\partial \beta_j}$$
$$\frac{l_i}{\theta_i} = \frac{y_i - b'(\theta_i)}{a(\phi)} = \frac{y_i - \mu_i}{a(\phi)}$$
$$\frac{\partial \mu_i}{\partial \theta_i} = \frac{\partial b'(\theta_i)}{\partial \theta_i} = b''(\theta_i) = \frac{\text{Var}[Y_i]}{a(\phi)}$$
$$\frac{\eta_i}{\beta_j} = z_{ij}$$

$\frac{\mu_i}{\eta_i}$ dipende dallaa g (link function) $\quad \mu_i = g^{-1}(\eta_i)$

$$\implies \frac{\partial l_i}{\partial \beta_j} = \frac{y_i - \mu_i}{a(\phi)} \cdot \frac{a(\phi)}{\text{Var}\left[Y_i\right]} \cdot \frac{\partial \mu_i}{\partial \eta_i} \cdot z_{ij} = \frac{(y_i - \mu_i) \cdot z_{ij}}{\text{Var}\left[Y_i\right]} \cdot \frac{\partial \mu_i}{\partial \eta_i}$$

$$\sum_{i=1}^{n} \left(\frac{y_i - \mu_i}{\text{Var}\left[Y_i\right]} \right) \cdot z_{ij} \cdot \frac{\partial \mu_i}{\partial \eta_i} = 0 \quad \forall j = 0...r$$

Applichiamo alle Bernuolli e Poisson

$$Y_i \sim Be(p_i) \quad \text{conn } \mu_i = p_i$$

$$\sum_{i=1}^{n} \left(\frac{y_i - p_i}{p_i(1-p_i)} \right) \cdot z_{ij} \cdot \frac{\partial p_i}{\partial \eta_i} = 0 \quad \forall j = 0...r$$

$$logit(p_i) = \eta_i \implies \text{invertendo la } logit() \;\; p_i = \frac{exp(\eta_i)}{1 + exp(\eta_i)}$$

$$\frac{\partial p_i}{\partial \eta_i} = \frac{exp(\eta_i)}{(1 + exp(\eta_i))^2} = \frac{exp(\eta_i)}{1 + exp(\eta_i)} = p_i(1 - p_i)$$

$$\sum_{i=1}^{n} \left(\frac{y_i - p_i}{\cancel{p_i(1-p_i)}} \right) \cdot z_{ij} \cdot \cancel{p_i(1-p_i)} = 0 \quad \forall j = 0...r$$

$$\sum_{i=1}^{n} \left(y_i - \frac{exp\left(\sum_j \beta_j z_{ij}\right)}{1 + exp\left(\sum_j \beta_j z_{ij}\right)} \right) = 0 \quad \forall j = 0..r$$

È difficile trovare le soluzioni in forma chiusa, quindi dovrò trovare le soluzioni con un metodo numerico.
I $\widehat{\beta}_{ML}$ sono le soluzioni numeriche di queste equazioni

$$Y_i \sim P(\lambda_i) \quad \text{con } \mu_i = \lambda_i$$

$$log(\mu_i) = \eta_i \implies \mu_i = exp(\eta_i) \implies \frac{\partial \mu_i}{\partial \eta_i} = exp(\eta_i) = \mu_i$$

$$\sum_{i=1}^{n} \frac{y_i - \mu_i}{\cancel{\text{Var}\left[y_i\right]}} \cdot z_{ij} \cdot \cancel{\mu_i} = \sum_{i=1}^{n} \left(y_i - exp(\sum_j \beta_j z_{ij}) \right) \cdot z_{ij} = 0 \quad \forall j = 0...r$$

Anche qui cercherò le soluzioni con metodi numerici

Ora devo ricordarmi che gli stimatori ML $\widehat{\beta}_{ML}$ sono asintoticamente gaussiani e asintoticamente efficienti
Per questo motivo, in R, nell'output, per esempio del modelli logistici, compare lo $z - value$

Lo $z - value$ è dato test $\begin{cases} H_0 : \beta_j = 0 \\ H_1 : \beta_j \neq 0 \end{cases}$ su cui viene usata la statistica $\dfrac{\widehat{\beta}_j}{SE(\widehat{\beta}_j)}$
Viene chiamato Z perché sotto H_0 questa statistica è asintoticamente gaussiana

Vediamo come trova lo standard error

Abbiamo $Cov(\widehat{\vec{\beta}}_{ML}) = \left(\mathbb{Z}^T \widehat{W} \mathbb{Z}\right)^{-1}$ dove W è una martrice diagonale con $W_{ii} = \dfrac{\left(\dfrac{\partial \mu_i}{\partial \eta_i}\right)}{\text{Var}\left[Y_i\right]}$ e $\widehat{W}$ è W calcolata facendo plug in dei $\widehat{\beta}$

Per le Bernoulli abbiamo $Cov(\widehat{\vec{\beta}}) = \left(\mathbb{Z}^T \, diag(\widehat{p}_i(1-\widehat{p}_i))\mathbb{Z}\right)^{-1}$ con $\widehat{p}_i = \dfrac{exp\left(\sum_j z_{ij}\widehat{\beta}_j\right)}{1 + esp\left(\sum_j z_{ij}\widehat{\beta}_j\right)}$

Per le Poisson abbiamo $W_{ii} = \mu_i \quad Cov(\widehat{\vec{\beta}}) = \left(\mathbb{Z}^T diag(\widehat{\mu})\mathbb{Z}\right)^{-1} \quad \widehat{\mu}_i = esp\left(\sum_j \widehat{\beta}_j z_{ij}\right)$

Con questi risultati potremo fare solo inferenza asintotica

Supponiamo di avere dei modelli annidati ovvero dei sotto modello M_0 di modelli completi M_1 a cui abbiamo tolto covariate
Quindi in generale $l(\widehat{\vec{\mu}}_0, \vec{y}) \leq l(\widehat{\vec{\mu}}_1, \vec{y})$
Vediamo l'ordinamento della devianza $= -2\big[l(\widehat{\vec{\mu}}, \vec{y}) - l(\vec{y}, \vec{y})\big]$ dei modelli annidati

$$D(\widehat{\vec{\mu}}_1, \vec{y}) = -2\big[l(\widehat{\vec{\mu}_1}, \vec{y}) - l(\vec{y}, \vec{y})\big] \leq -2\big[l(\widehat{\vec{\mu}_0}, \vec{y}) - l(\vec{y}, \vec{y})\big] = D(\widehat{\vec{\mu}}_0, \vec{y})$$

Vogliamo valutare questa perdita di devianza e capirne l'entità

$$D(\widehat{\vec{\mu}}_1, \vec{y}) - D(\widehat{\vec{\mu}}_0, \vec{y}) = -2\big[l(\widehat{\vec{\mu}_1}, \vec{y}) - l(\widehat{\vec{\mu}}_0, \vec{y})\big] = -2log\left(\widehat{\lambda}\right) = \frac{L(\widehat{\vec{\mu}}_0, \vec{y})}{L(\widehat{\vec{\mu}}_1, \vec{y})}$$

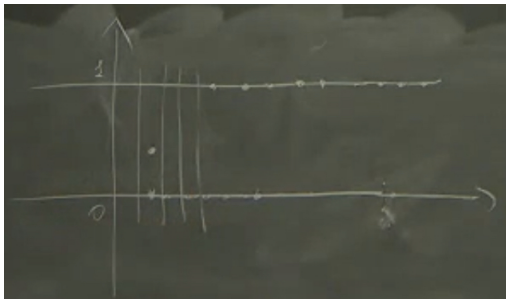
Faremo un test asintotico su $-2log(\lambda)$, ovvero un test χ^2
Con $p - value$ alti i modelli sono equivalenti, se invece $p - value$ bassi allora i modelli non sono equivalenti

9.4 Regressione logistica semplice

$$p_i = \frac{exp\left\{\sum_j \beta_j z_{ij}\right\}}{1 + exp\left\{\sum_j \beta_j z_{ij}\right\}}$$

Supponiamo di avere delle osservazioni (z_i, y_i), che saranno 0 o 1 essendo per un sistema generalizzato, vogliamo stimare la media al variare di z_i

Se avessimo tantissime osservazioni basterebbe fissare z_i e fare la media. Questo in generale non si fa, suddivideremo in bande la covariata e valuteremo la media in ognuna di queste



Se $\beta_1 > 0$ allora al crescere della covariata Z_1, cresce la probabilità di successo $(y = 1)$

Se invece $\beta_1 < 0$ allora al crescere della covariata Z_1, decresce la probabilità di una risposta 1

Oss. In questo modello è più difficile interpretare quantitativamente il β, però il parametro non parametrizza proprio la variazione della probabilità, ma più che altro l'ODDS

Data una probabilità p, $ODDS = \dfrac{p}{1-p}$

Date due probabilità p_1, p_2, l'Odds Ratio è il rapporto tra gli odds $\dfrac{\frac{p_1}{1-p_1}}{\frac{p_2}{1-p_2}}$

Cerchiamo una relazione tra Odds Ratio e i β_j

Supponiamo di avere Z continua e di avere $p = \dfrac{exp(\eta)}{1 - exp(\eta)}$

Vogliamo confrontare $p(z+1)$ e $p(z)$:

$$ODDS = \frac{\frac{exp(\beta_0 + \beta_1(Z+1))}{1 + exp(\beta_0 + \beta_1(Z+1))}}{\frac{1}{1 + exp(\beta_0 + \beta_1(Z+1))}} = exp(\beta_0 + \beta_1(Z+1))$$

$$\implies ODDS\ RATIO = OR = \frac{exp(\beta_0 + \beta_1(Z+1))}{exp(\beta_0 + \beta_1 Z)} = exp(\beta_1)$$

Quindi OR nell'incrementare di 1 la covariata continua è e^{β_j}, se invece la covariata fosse categorica avremmo OR per passare da una categoria alla successiva è e^{β_j}

Diremo che e^{β_j} è l'OR relativo alla Z_j

Possiamo costruire un IC per l'OR

IC asintotico di livello $1 - \alpha$ per β_j è $[\widehat{\beta}_j \pm z_{1-\frac{\alpha}{2}}\ se(\widehat{\beta}_j)]$

$\implies$ IC asintotico di livello $1-\alpha$ per e^{β_j} è $[exp\{\widehat{\beta}_j \pm z_{1-\frac{\alpha}{2}}\, se(\widehat{\beta}_j)\}]$

Per poter usare questo modello serve "almeno" per ogni predittore 10 zeri e 10 uni, perché se i dati sono molto sbilanciati, ovvero prevalenza di zeri o uni, in questo caso è difficile far previsione

Anche qui per confrontare due modelli diversi si usa AIC

Esempio di indici di Goodness of fit è Breierscore $= \frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{p}_i)$, questa mi dice quanto dista l'osservazione dalla probabilità stimata

Lezione 27 (24/05/23)

9.5 Regressione logistica come classificatore

A partire dalle stime di θ_i vogliamo prevedere le stime $\widehat{y}_i \in \{0,1\}$
Nel pratico troveremo un valore di soglia $p_0 = \frac{1}{2}$, se $\widehat{p}_i \geq p_0 \implies \widehat{y}_i = 1$ invece se $\widehat{p}_i \leq p_0 \implies \widehat{y}_i = 0$

Valutiamo la bontà di queste stime

$\widehat{y}, y \in \{0,1\}$ quindi divido in quattro casi:

$\widehat{y} \setminus y$	0	1
0	n_{00}	n_{01}
1	n_{10}	n_{11}

Nei casi n_{00} e n_{11} la stima è giusta, invece in n_{01} e n_{10} ci sono errori:
Errore di misclassificazione $= \frac{n_{01}+n_{10}}{n}$
Rate dei corretti classificati $= \frac{n_{00}+n_{11}}{n}$

Sensibilità $= \mathbb{P}(\widehat{y}=1|y=1) = \frac{n_{11}}{n_{n_{11}+n_{01}}}$
Specificità $= \mathbb{P}(\widehat{y}=0|y=0) = \frac{n_{00}}{n_{n_{00}+n_{10}}}$

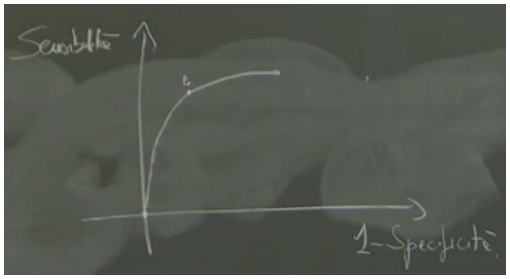
Oss. È più importante massimizzare la sensibilità, perché i falsi negativi sono più gravi dei falsi positivi

Curva ROC: receiver operating curve

$\forall p_0 \in [0,1]$ calcolo la sensibilità e la specificità

Esempio, per $p_0 = 0$ avrò sens=0 e spec=1, invece per $p_0 = 1$ avrò sen=1 e spe=0

Plotto (1-specificità)$\times$(sensibilità)



Indice AUC, area sotto la curva, preferirò classificatori con AUC massima, perché hanno sensibilità maggiore

Come p_0 andremo a scegliere il punto che massimizza la sensibilità e minimizza 1-specificità

Applichiamo il classificatore ad un test diagnostico che può avere un test positivo T_+ o negativo T_-

Si valuta la probabilità di essere effettivamente malati quando si riceve un test positivo

$$\mathbb{P}(M|T_+) = \frac{\mathbb{P}(T_+|M)\,\mathbb{P}(M)}{\mathbb{P}(T_+|M)\,\mathbb{P}(M) + \mathbb{P}\left(T|M^c\right)\mathbb{P}(M^c)}$$

Dove $\mathbb{P}(T_+|M)\,\mathbb{P}(M)$ è la sensibilità e $\mathbb{P}(T_+|M^c)\,\mathbb{P}(M^c)$ è la 1-specificità

10 Statistica non parametrica

Vediamo dei modelli che non sono parametrici, ovvero le variabili non sono note a meno di k parametri

10.1 Test d'indipendenza

Test validi per ogni tipo di variabile, che controllano se due variabili sono tra di loro indipendenti. Non facciamo assunzioni sulle leggi delle variabili aleatorie.

Prendiamo A,B due variabili categoriche o comunque discrete

$A \setminus B$	1	2	
1	n_{11}	n_{12}	$n_{1\cdot}$
2	n_{21}	n_{22}	$n_{2\cdot}$
	$n_{\cdot 1}$	$n_{\cdot 2}$	$n_{\cdot\cdot}$

n_{ij} = frequenze osservate O_{ij}
E_{ij} = frequenze attese = $\dfrac{n_{i\cdot}n_{\cdot j}}{n_{\cdot\cdot}}$

Oss. Se fossero indipendenti le frequenze sarebbero il prodotto delle frequenze marginali

Posto il test $\begin{cases} H_0 : A \perp\!\!\!\perp B \\ H_1 : A \text{ dipendente } B \end{cases}$

Statistica test = $\sum_j \dfrac{(O_{iij} - Eij)^2}{E_{ij}} \sim \chi^2((r-1)(c-1))$ dove $r = \#$ righe e $c = \#$ colonne

$$RC\big(\chi^2 > \chi^2_{1-\alpha}((r-1)(c-1))\big)$$

Oss. Nell'esempio di A e B 2×2 il test χ^2 coincide coincide con il confronto tra proporzioni

Oss. È possibile che ci sia 0 nella tabella, per cui è difficile che sia indipendente, per affrontare situazioni "problematiche" ci sono diverse correzioni al test χ^2

10.2 Test di Buon adattamento

Stiamo verificando che la variabile osservata è distribuita come una variabile nota

Per variabili aleatorie discrete parametriche $x_1 ... x_n$ $H_0 : F \sim P(\lambda)$

Per variabili aleatorie continue parametriche, per affrontare queste dovrò parametrizzare in intervalli

Ci sono molti altri test non parametrici, per esempio shapiro test o Anderson-Darling...

Test di Markov che si basano sul teorema di Glivenko-Cantelli

Voglio verificare che la mia variabile abbia una certa distribuzione $\begin{cases} H_0 : F \sim F_0 \\ H_1 : F \not\sim F_0 \end{cases}$

Questi test usano la statistica $D = \underset{x}{sup}|\widehat{F}_n(x) - F_0(x)|$ si cerca la distribuzione di D sotto H_0 e per teo G.C. il valore D tende a zero, quindi per valori grandi di D, sarò portato a rifiutare H_0.

Dato che D sotto H_0 ha distribuzione nota, posso usarlo per ogni test di adattamento

10.3 Confronto tra distribuzioni non gaussiane

Test Wilcoxson è un test di confronto tra distribuzioni non gaussiane e indipendenti

Vediamo un esempio, senza spiegare il metodo precisamente

Si studiano i tempi di rottura di un sistema per valutare il miglior tipo di manutenzione

Abbiamo osservazioni dei tempi di rottura per due tipi di manutenzione I e II

I	7	26	10	8	29
II	3	150	40	34	32

Ci sono poche osservazioni, quindi è improbabile siano gaussiane, non possiamo procedere con i metodi noti

$\begin{cases} H_0 : F_I(t) = F_{II}(t) \\ H_1 : \text{le due fdr sono diverse e stocasticamente ordinate} \end{cases}$

Mi dimentico delle osservazione e costruisco i ranghi, ovvero la posizione del dato una volta ordinati

I	2	5	4	3	6
II	1	10	9	8	7

L'idea è che se i tempi arrivassero dalla stessa popolazione, allora mi aspetterei che i ranghi siano mescolati, quindi la somma dei ranghi dei due tipi dovrebbe essere simile

Sotto H_0 $\mathbb{P}(R_1 = r_1; R_2 = r_2 ... R_n = r_n) = \dfrac{1}{n!}$ perché supponendo che non ci sia differenza (H_0) allora tutte le stringhe ha la stessa probabilità

Chiamata $W = \sum_{II} R_j$ e calcolato $\sum_{II} R_j = 35$

Si può calcolare, usando il calcolo combinatorio, $\mathbb{P}(W \geq 35) = 0.07654$ questo equivale al $p - value$ perché è la probabilità di una statistica rispetto al valore osservato

Quindi in questo caso c'è un pochino di evidenza per dire che i due gruppi hanno distribuzioni diverse

Test di Mann-Whitney

Serve per due popolazioni accoppiate $x_1...x_n \;\; y_1...y_n$

Nel caso parametrico si usava il Paired t-test e quindi $d_i = y_i - x_i$ e la statistica test in funzione di d_i

Nel caso non parametrico uso comunque d_i, ma il test diventa: $H_0 : F_d = F_{-d}$

In H_0 si pone $F_d = \overline{F}_d = 1 - F_d$

Esempio: Si raccolgono dati sulle miglia per gallone per una macchina prima e dopo l'aggiunta di additivo

Si vuole dimostrare che dopo l'aggiunta di additivo le miglia per gallone salgono

Macchina	1	2	3	4	5	6	7
Prima x	17.2	21.6	19.5	19.1	22.0	18.7	20.3
Dopo y	18.3	20.8	20.9	21.2	22.7	18.6	21.9
$y - x$	1.1	-0.8	1.4	2.1	0.7	-0.1	1.6
Rango con segno	4	-3	5	7	2	-1	6

Dove i ranghi sono calcolati rispetto a $|y - x|$

Sommiamo i ranghi positivi $\sum T_i = 24$ anche qui se le popolazioni sono uguali e quindi la differenza è simmetrica, allora la somma dei ranghi positivi e negativi saranno opposte, ma simili e quindi posso calcolare $p - value = 0.0547$, perciò c'è debole evidenza statistica per cui la distribuzione dopo sia maggiore della distribuzione prima

Nel caso di più variabili c'è l'estensione di questo test che si chiama test di Kruska-Wallis