STATISTICA INFERENZIALE

Scopo — Ottenere info sulla distribuzione di una popolazione studiando dati rilevanti su un campione.

Statistica inferenziale — La popolazione non è nota, prendo un campione e tramite l'analisi dico qualcosa sulla popolazione in modo INDUTTIVO.

Ipotesi chiave — a) ogni nuova estrazione è indipendente dalle precedenti → indipendenza
b) l'estrazione di ogni individuo che andrà a costruire il campione è fatta in modo identico al precedente → equidistribuzione

Situazione — (Ω, P) spazio di probabilità su cui sono definite tutte le v.a. che considereremo.

- Il carattere che vogliamo studiare è rappresentato da una v.a. $X: \Omega \to \mathbb{R}$ con distribuzione P_X sconosciuta (in parte o totalmente);
- per determinare P_X, suppongo di avere n $X_1, \dots, X_n$ v.a. i.i.d. che rappresentano il campione estratto.

Campione statistico — Si intende di taglia n per X il dato di $X_1, \dots, X_n$ con X_i v.a. i.i.d. con legge P_X.

es. altezza $\sim N(m, \sigma^2)$ mi aspetto che sia distribuita così ma non sono conosciuti i parametri. Cerco delle stime sui parametri tramite la STATISTICA PARAMETRICA.

Statistica parametrica — Scopo:
- stima parametrica
- intervalli di fiducia
- test statistici

Notazione:
- β: indica un parametro non noto della popolazione;
- $\beta \in H$: H indica l'insieme dei valori ammissibili;
- P_X^{β}: la β indica che la prob. dipendente da un parametro.

Statistica campionaria — Funzione $g(X_1, \dots, X_n)$ dal campione $(X_1, \dots, X_n)$

Stimatori di un parametro β — Una statistica campionaria $P(X_1, \dots, X_n)$ atta a stimare β

es. Per stimare l'altezza di tutti gli iscritti alla triennale di informatica è $\overline{X}_{100} = \frac{X_1 + \dots + X_{100}}{100}$

Stimatore corretto (non distorto)

$g(X_1,\dots,X_n)$ stimatore del parametro β si dice corretto se $E[g(X_1,\dots,X_n)]=\beta$

Proposizione

$X_1,\dots,X_n$ campione statistico con momento secondo, allora:
① $\overline{X}_n$ è uno stimatore corretto per $E[X_i]=E[X]$;
② S_n^2 è uno stimatore corretto per $Var(X_i)=Var(X)$

Dimostrazione:
① $\overline{X}_n$ stimatore corretto per $E[X_i] \Leftrightarrow E[\overline{X}_n]=E[X]$
$E\left[\frac{X_1+\dots+X_n}{n}\right]=\frac{1}{n}\sum_{i=1}^{n}X_i=\frac{1}{n}\cdot nE[X]=E[X]$

② S_n^2 stimatore corretto per $Var(X) \Leftrightarrow E[S_n^2]=Var(X)$
DA FARE DIM.

Stimatore Consistente

Parametro β della distribuzione, $X_1,\dots,X_n$ $n\in\mathbb{N}$ campione di ∞ v.a. i.i.d. di X
$g_n(X_1,\dots,X_n)$ è uno stimatore consistente di β se per $n\to\infty$ vale $g_n(X_1,\dots,X_n)\to\beta$ in probabilità

es. $X_1,\dots,X_n$ campione di v.a. i.i.d. con momento secondo
Allora $\overline{X}_n$ è uno stimatore consistente per $E[X_i]$ $(=E[X]=\mu)$
e S_n^2 è uno stimatore consistente per $var(X_i)$

Dim. si dimostra con LGN

Efficienza di uno stimatore

Presi $g(X_1,\dots,X_n)$ e $h(X_1,\dots,X_n)$ stimatori corretti del parametro β dico che g è più efficiente di h se $var(g(X_1,\dots,X_n))\le var(h(X_1,\dots,X_n))$

es. $X_1,\dots,X_n$ v.a. i.i.d. $\Rightarrow$ $\overline{X}_n$ è tanto più efficiente quanto più n aumenta

Dim. $Var(\overline{X}_n)=var\left(\frac{X_1+\dots+X_n}{n}\right)=\frac{1}{n^2}\cdot(var(X_1)+\dots+var(X_n))=\frac{1}{n^2}\cdot n\,var(X_i)=\frac{var(X_i)}{n}$
quindi più cresce n e più diventa piccolo il risultato, si disperde meno

Metodo della massima verosimiglianza

Metodo per scegliere un buon stimatore. Ho $(X_1,\dots,X_n)$ campione i.i.d di X, X con distribuzione P_X^β con β parametro.
Le v.a. possono essere:
- discrete con funzione di massa $P_\beta(x)$
- con densità $f_\beta(x)$

Funzione di verosimiglianza

$L_\beta:\mathbb{R}^n\to\mathbb{R}$ definita come:
- caso discreto: $P_\beta(X_1,\dots,X_n)=P\{X_1=x_1\}\cdot\ldots\cdot P\{X_n=x_n\}=P\{X_1=x_1,\dots,X_n=x_n\}$
- caso con densità: $f_\beta(X_1)\cdot\ldots\cdot f_\beta(X_n)$

Stima massima verosimiglianza (MLE)

Si chiama MLE se $\exists$ una statistica campionaria $\hat{\beta}=\hat{\beta}(X_1,\dots,X_n)$ tale che valga l'uguaglianza $L_{\hat{\beta}}(X_1,\dots,X_n)=\max_{\beta\in H}L_\beta(X_1,\dots,X_n)$ $\forall(X_1,\dots,X_n)\in\mathbb{R}^n$

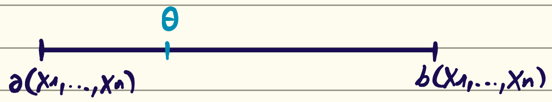
Il massimo si trova tramite la derivata della funzione
Ho $y=L(\beta)$ e cerco $\frac{dL}{d\beta}=(\hat{\beta})$
Alle volte non basta una stima ma servono degli intervalli di fiducia

Intervalli di fiducia

$(X_1,\dots,X_n)$ campione statistico di legge P_θ, $\theta \in H \in \mathbb{R}$
$\alpha \in (0,1)$ e $a,b:\mathbb{R}^n \to \mathbb{R}$ funzioni t.c. $a(X_1,\dots,X_n)$, $b(X_1,\dots,X_n)$ siano v.a.

Allora un intervallo aleatorio $I[a(X_1,\dots,X_n), b(X_1,\dots,X_n)]$ si dice intervallo di fiducia per θ di livello $1-\alpha$ se

$$\forall \theta \in H \quad P_\theta(\theta \in I) = P(a(X_1,\dots,X_n) \le \theta,\ b(X_1,\dots,X_n) \ge \theta) \ge 1-\alpha$$



La prob. che θ stia in questo intervallo è $\ge 1-\alpha$