9 Stimatori e loro proprietà

Definition 78. *Dediniamo campione aleatorio (random sample) un insieme di variabili aleatorie indipendenti ed identicamente distribuite $X_1, ..., X_n$ definite su uno spazio di probabilità $\{\Omega, \mathfrak{S}, \mathbb{P}\}$.*

Spiegazione Semplice

Cos'è un "Campione Aleatorio"?
Questa è la definizione base della statistica. Un "campione aleatorio" è un insieme di n osservazioni (i tuoi dati, $X_1, \ldots, X_n$) che rispetta due condizioni cruciali:

- **Indipendenti (I):** Il valore di un'osservazione non influenza le altre. (Es. L'esito del primo lancio di un dado non ha effetto sul secondo lancio).
- **Identicamente Distribuite (ID):** Tutte le osservazioni provengono dalla stessa "fonte", cioè seguono la stessa identica distribuzione di probabilità. (Es. Tutti i lanci provengono dallo *stesso* dado, non da dadi truccati diversi).

Queste due condizioni insieme sono abbreviate in **i.i.d.** e sono l'ipotesi di partenza per la maggior parte dei teoremi statistici.

Remark 79. *In realtà né l'ipotesi di indipendenza né quella di identica distribuzione sono necessarie e saranno entrambe abbandonate più avanti; partiamo da queste condizioni super-semplificate per necessità.*

Supponiamo ora che la misura di probabilità P sia nota solo a meno del valore di un certo parametro θ. In genere, si definisce Statistica Parametrica la disciplina che studia le tecniche per risalire al valore di θ nel caso in cui esso appartenga ad uno spazio di dimensione finita, $\theta \in \mathbb{R}^p$, $p \in \mathbb{N}$; si parla invece di Statistica NonParametrica la disciplina che studia il caso in cui θ ha dimensione infinita. Ad esempio, se sappiamo che P è una legge Gaussiana ma ne ignoriamo valor medio e varianza siamo nell'ambito della Statistica Parametrica; se sappiamo che P ammette una densità e cerchiamo di ricavarla senza sapere che tipo di legge sia, siamo nell'ambito della Statistica Nonparametrica. Per distinguere un valore generico del parametro dal valore "vero" che caratterizza la legge da cui è stato estratto un certo campione aleatorio, puòà essere conveniente scrivere θ_0 in questo ultimo caso.

Spiegazione Semplice

Statistica Parametrica vs. Nonparametrica
Questo passaggio definisce i due rami principali della statistica:

- **Statistica Parametrica:** È come giocare a un indovinello in cui *conosci già le regole*. Tu *ipotizzi* che i tuoi dati seguano una forma specifica (es. una curva a campana Gaussiana), ma non conosci i "parametri" che la definiscono (es. la media μ o la varianza σ^2). Il tuo obiettivo è "indovinare" (stimare) questi parametri.
- **Statistica Nonparametrica:** È come giocare a un indovinello in cui *non conosci le regole*. Non fai nessuna ipotesi sulla forma della distribuzione da cui provengono i dati. Il tuo obiettivo è molto più difficile: cercare di "ricostruire" la forma stessa della distribuzione (che è un oggetto a "dimensione infinita").

Definition 80 (Stimatore)**.** *Uno stimatore di un certo parametro $\theta \in \mathbb{R}^p$ è una funzione (o meglio una successione di funzioni) misurabile $T_n : \mathbb{R}^n \to \mathbb{R}^p$ di un campione aleatorio; scriveremo equivalentemente*

$$T_n = \hat{\theta}_n = T_n(X_1, ..., X_n).$$

Spiegazione Semplice

Cos'è uno "Stimatore"?
Uno stimatore (indicato con $\hat{\theta}_n$, letto "theta-cappello") è semplicemente una **ricetta** (una formula

matematica) che usa i dati del tuo campione $(X_1, \ldots, X_n)$ per produrre una "stima", cioè un tentativo di indovinare il valore del parametro vero θ.

- **Parametro θ:** Il valore *vero* e ignoto nella popolazione (es. la vera altezza media di tutti gli italiani).
- **Stimatore T_n:** La *ricetta* che usi (es. la formula della media aritmetica).
- **Stima $\hat{\theta}_n$:** Il *numero* che ottieni applicando la ricetta al tuo campione (es. 175cm, la media calcolata sul tuo campione di 100 persone).

Remark 81. *La definizione è volutamente molto generica: qualsiasi funzione misurabile può essere considerata uno stimatore, quindi il punto cruciale sarà determinare quali siano le proprietà che consentono di valutare la bontà della scelta fatta.*

Example 82. *L'esempio più ovvio di stimatore (per il valore medio $\mu = E[X_i]$) è la media aritmetica, cioè*

$$\tilde{X}_n = \tilde{\mu}_n := \frac{1}{n}\sum_{i=1}^{n} X_i$$

Altro stimatore naturale è la cosiddetta varianza empirica (che stima $\sigma^2 = Var[X_i]$), cioè

$$S_n^2 = \tilde{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \tilde{\mu}_n)^2.$$

9.1 Criteri per Valutare gli stimatori

- Non-distorsione/asintotica non distorsione
- Consistenza (in probabilità, in media r-esima, quasi certa, completa)
- Efficienza (assoluta e relativa)
- Asintotica Gaussianità

Spiegazione Semplice

La "Pagella" di uno Stimatore
Dato che "qualsiasi formula" può essere uno stimatore (come dice il Remark 81), come facciamo a dire se uno stimatore è "buono"? Controlliamo queste quattro proprietà.

1. **Non-distorsione (Unbiasedness):** Lo stimatore è "onesto"? In media, ci azzecca, o tende a sbagliare sempre un po' troppo in alto o troppo in basso? Se $\mathbb{E}[\hat{\theta}_n] = \theta$, è non-distorto (onesto). Se no, $\mathbb{E}[\hat{\theta}_n] - \theta$ è il suo "bias" (distorsione).
2. **Consistenza (Consistency):** Lo stimatore "impara" dai dati? Se aumentiamo la dimensione del campione ($n \to \infty$), la stima $\hat{\theta}_n$ converge al valore vero θ? (Questa è la proprietà fondamentale. Se non è consistente, è inutile).
3. **Efficienza (Efficiency):** Lo stimatore è "preciso"? Tra due stimatori onesti (non distorti), preferiamo quello che ha la varianza più bassa, cioè quello i cui valori "ballano" di meno attorno al valore vero.
4. **Asintotica Gaussianità (Asymptotic Normality):** Per campioni grandi (n grande), la distribuzione degli errori dello stimatore $\hat{\theta}_n$ assomiglia a una curva a campana (Gaussiana)? Questo è fondamentale perché ci permette di usare il Teorema del Limite Centrale per costruire intervalli di confidenza (es. "siamo al 95% sicuri che il valore vero sia...").

Example 83. *Media aritmetica banale mostrare non-distorsione, consistenza ed asintotica Gaussianità; si può mostrare con Cramer-Rao (vedi dopo) che efficiente in senso assoluto nel caso Gaussiano, nel caso non-Gaussiano, si consideri un altro stimatore lineare*

$$\sum_{i=1}^{n} c_i X_i, \sum_{i=1}^{n} c_i = 1$$

(dove la condizione che la somma dei pesi sia 1 garantisce la non-distorsione). La varianza è evidentemente $\sum_{i=1}^{n} c_i^2 Var[X_i]$*, e pertanto nel caso di variabili incorrelate ed identicamente distribuite possiamo porcil problema di massimo vincolato*

$$\min_{c_i} \sum_{i=1}^{n} c_i^2 \quad s.t. \sum_{i=1}^{n} c_i = 1.$$

Con i moltiplicatori di Lagrange troviamo

$$\psi(c_1, ..., c_n; \lambda) = \sum_{i=1}^{n} c_i^2 + \lambda\{\sum_{i=1}^{n} c_i - 1\}$$

con derivate prime

$$\frac{\partial \psi(c_1, ..., c_n; \lambda)}{\partial c_i} = 2c_i + \lambda.$$

Ne segue che il problema di minimo sarà risolto imponendo che i coefficienti c_i *siano tutti uguali e pertanto pari a* n^{-1}*; questo rende la media aritmetica uno stimatore BLUE (Best Linear Unbiased Estimator).*

Spiegazione Semplice

Perché la media aritmetica è così speciale? (BLUE)
Questo esempio dimostra che la media aritmetica ($\overline{X}$) ha un'ottima pagella. In particolare, è il **BLUE**: *Best Linear Unbiased Estimator* (Miglior Stimatore Lineare Non Distorto).

- **Linear (Lineare):** È una combinazione lineare (una somma pesata) dei dati: $c_1X_1 + \cdots + c_nX_n$.
- **Unbiased (Non Distorto):** Per essere non distorto, la somma dei pesi deve fare 1 ($\sum c_i = 1$).
- **Best (Migliore):** Tra *tutti* gli stimatori lineari non distorti, la media aritmetica (dove $c_i = 1/n$ per tutti) è quello che ha la varianza più piccola in assoluto.

La dimostrazione nel testo usa l'analisi matematica (i moltiplicatori di Lagrange) per trovare i c_i che minimizzano la varianza, con il vincolo che la loro somma sia 1. Il risultato è $c_i = 1/n$.

Example 84 (Varianza Campionaria)**.** *In questo caso possiamo scrivere*

$$\begin{aligned} S_n^2 &= \frac{1}{n}\sum_{i=1}^{n}(X_i - \tilde{\mu}_n)^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \tilde{\mu}_n^2 \\ E[S_n^2] &= \frac{1}{n}\sum_{i=1}^{n} E[X_i^2] - (E[\tilde{\mu}_n])^2 - Var[\tilde{\mu}_n] \\ &= \frac{1}{n}\sum_{i=1}^{n}(\sigma^2 + \mu^2) - \mu^2 - \frac{\sigma^2}{n} \\ &= (\sigma^2 + \mu^2) - \mu^2 - \frac{\sigma^2}{n} = \sigma^2 - \frac{\sigma^2}{n} = \frac{n-1}{n}\sigma^2. \end{aligned}$$

La varianza campionaria risulta quindi distorta, anche se asintoticamente non- distorta; in un certo senso è come se perdessimo un grado di liberta centrando le osservazioni sulla media empirica invece di quella teorica (per $n = 1$ *la varianza campionaria è identicamente nulla). Per quello che riguarda la consistenza debole, è una semplice conseguenza del Lemma di Slutzky; se le* X_i *hanno momenti quarti finiti, si può dimostrare la convergenza in media quadratica usando la disuguaglianza di Chebyshev.*

Spiegazione Semplice

Il problema della Varianza Campionaria (Bias)
Questo è un risultato classico e importante.

- **Lo stimatore S_n^2:** La formula intuitiva per la varianza, dove si divide per n.
- **Il risultato $E[S_n^2] = \frac{n-1}{n}\sigma^2$:** La media di questo stimatore *non* è σ^2 (il valore vero), ma è σ^2 moltiplicato per $\frac{n-1}{n}$ (un numero leggermente più piccolo di 1).
- **Conclusione: S_n^2 è distorto (biased).** Tende a sottostimare *leggermente* la vera varianza.

Perché? Come dice il testo, "perdiamo un grado di libertà". Stiamo usando la media campionaria $\tilde{\mu}_n$ (calcolata dai dati) al posto della vera media μ (che non conosciamo). I dati sono, in media, più vicini alla *loro* media $\tilde{\mu}_n$ di quanto non lo siano alla media *vera* μ. Questo fa sì che la somma dei quadrati degli scarti sia leggermente più piccola di quanto dovrebbe essere.
La soluzione: Per "correggere" questa distorsione, gli statistici usano la *varianza campionaria corretta*, S_{n-1}^2, che divide per $n-1$ invece che per n.
Asintoticamente non distorta: Per n grande (es. $n = 1000$), il fattore $\frac{999}{1000}$ è quasi 1. Quindi, anche se S_n^2 è tecnicamente distorto, il suo bias (errore) svanisce man mano che il campione cresce.

Exercise 85. *Determinare opportune condizioni per dimostrare la convergenza quasi certa della varianza campionaria al valore della varianza σ^2; sotto quali condizioni vale il teorema del limte centrale?*