CHAPTER 4

Stime di Parametri

4.1 La statistica

La statistica studia il problema di inferire da un campione i parametri e/o i modelli che descrivono la popolazione dalla quale il campione è stato estratto. In particolare possiamo dividerla in due categorie:

- Stima dei parametri, misura di una quantità fisica;
- Test d'ipotesi, ovvero la prova della validità di un modello.

Una funzione dipendente da N misure di un campione $f(x_1, ..., x_n)$ si chiama **statistica**, essa è una **variabile aleatoria**. Quindi segue una sua distribuzione di probabilità pdf_f derivabile dalla joint-pdf dei campionamenti e dalla forma della funzione f.
Complessivamente si hanno 3 pdf:

- la pdf_x(x,θ) delle singole misure campionate;
- la $pdf_{set}(x_1, ..., x_n, \theta)$ dei campionamenti (che è multidimensionale);
- la pdf_f della statistica dei campionamenti (dipende dalla forma funzionale di f).

4.2 Stimatori

Sia data una p.d.f. (probability distribution function), $f(x, \theta)$ di una variabile x aleatoria continua e dipendente da un parametro θ, di cui non conos-

ciamo il vero valore θ_{true}.
Se si possiede un insieme $\{x_i\}_i^N$ di N misure della variabile x, possiamo chiederci se sia possibile determinare una stima del parametro θ_{true} in funzione di tali misure, $\hat{\theta} = \hat{\theta}(x_1, ..., x_N)$, le funzioni di questo tipo prendono il nome di **stimatori**. Uno stimatore è una statistica opportunamente scelta. Con **stima** s'intende il valore assunto $\hat{\theta}^*$ dallo stimatore per uno specifico campione.

Figure 4.1: N misure

Poichè lo stimatore $\hat{\theta}$ è dipende da variabili aleatorie è anch'esso una variabile aleatoria e dunque si può parlare di valore medio $E[\hat{\theta}]$ e varianza $V[\hat{\theta}]$ di una particolare stima, oltre ad avere una sua pdf($\hat{\theta}$).

$$\hat{\theta} \pm \sigma_\theta \tag{4.1}$$

Di conseguenza un insieme di misure restituirà un solo valore appartenente ad una popolazione ottenuta da campione fatto di misure della variabile aleatoria presa in considerazione.

4.3 Proprietà degli stimatori

Consideriamo un campione di N misure $\{x_i\}_i^N$ vengono definite IID (Independent Identically Distributed) quando sono:

- **indipendenti:** l'esito di una misura non è influenzato dalle misure precedenti;
- **identiche:** Delle misure vengono definite identiche quando tutte quante seguono la stessa distribuzione di probabilità

Nella statistica alle stime si possono associare diverse caratteristiche :

1. **Consistenza:** una stima si dice consistente quando all'aumentare del numero di misure (convergenza probabilistica) si converge al valore vero del parametro. Ossia quando:

$$\lim_{N \to \infty} \hat{\theta}(x_1, ..., x_N) = \theta_{true} \tag{4.2}$$

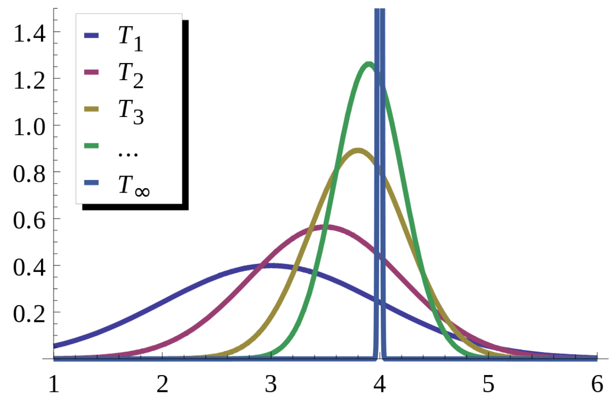


Figure 4.2: Proprietà di consistenza di uno stimatore

2. **Biased**:

   (a) una stima si dice unbiased o imparziale, se mediamente coincide con il valore vero del parametro, ovvero

   $$b_n(\hat{\theta}) = E(\hat{\theta}_n - \theta_{true}) = E(\hat{\theta}_n) - E(\theta_{true}) = 0 \iff E(\hat{\theta_n}) = \theta_{true} \tag{4.3}$$

   (b) Una stima si dice asintoticamente unibiased se $b_n(\hat{\theta}) \to 0$ per $n \to \infty$.

   Si osserva che $b_n(\hat{\theta})$ è uno stimatore lineare, dunque se $\hat{\theta}$ è stimatore di θ_{true} questo non vuol dire che $\hat{\theta}^2$ è stimatore di θ^2_{true}.

3. **Efficienza**: si dice che una stima è più efficiente di un'altra se la sua

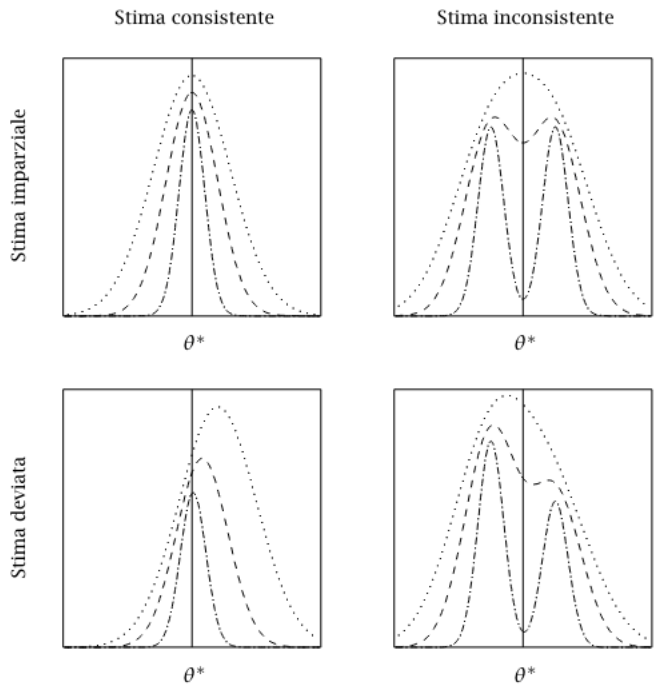


Figure 4.3: Consistenza e Bias di uno stimatore

varianza è inferiore, quindi se mediamente essa è più vicina al valore centrale $E(\hat{\theta})$, che coincide con θ_{true} se la stima è anche imparziale (ubiased).

4. **Varianza:** desideriamo che ripetendo i campionamenti le stime ottenute siano tutte vicine tra loro, ovvero la varianza della $pdf(\hat{\theta})$ sia il più piccola possibile.

4.3.1 Precisione e Accuratezza

Per uno stesso parametro si possono in generale definire tanti stimatori diversi tra loro, ma non tutti hanno le proprietà desiderate. Da notare che non è detto che esista (o sia possibile trovare) uno stimatore che soddisfi contemporaneamente tutte le proprietà richieste.

Esempio

La media delle misure è uno stimatore non distorto.

Dim.
Definito come stimatore la media aritmetica delle misure di un campione IID:

$$\hat{\mu} = \frac{1}{N}\sum x_i$$

si ha che il valore di aspettazione dello stimatore è:

$$E(\hat{\mu}) = E(\frac{1}{N}\sum x_i) = \frac{1}{N}\sum E(x_i) = \frac{1}{N}\cdot N \cdot \mu_t = \mu_t$$

quindi la media aritmetica è uno stimatore **non distorto** poichè:

$$b_n(\hat{\mu}) = E(\hat{\mu}) - \mu_t = \mu_t - \mu_t = 0$$

Se la pdf(x) delle misure soddisfa le ipotesi del TCL, la pdf($\hat{\mu}$) per $N \to \infty$ tende a una **gaussiana** con media μ e varianza $\frac{\sigma^2}{N}$ si ha che $\hat{\mu}$ è uno stimatore **consistente**. Poichè $V[\hat{\mu}] = \frac{\sigma^2}{N}$ al crescere del numero di campionamenti la varianza si riduce e dunque le stime ottenute con diversi set di dati sono tutte vicine tra loro.

4.3.2 Incertezze sulle stime

Uno stimatore come ogni altra variabile aleatoria è soggetto a due tipi d'incertezze:

1. **Incertezza sistematica:** nel caso di misure biased esiste una differenza sistematica fra la misura sperimentale ottenuta e il valor vero, ed è uguale per tutte le misure e non è possibile determinarlo essendo una proprietà intrinseca.

2. **Incertezza statistica:** è associata alla precisione, e la si può ridurre aumentando il numero di misure o cambiando l'apparato sperimentale.

$$MSE = E[(\hat{\theta} - \theta_t)^2] = Var(\hat{\theta}) \ + \ b_n^2(\hat{\theta}) \tag{4.4}$$

Definisce l'errore quadratico medio e tiene conto sia dell'errore statistico misurato dalla varianza che dell'errore sistematico misurato dal bias.

4.3.3 La Varianza come stimatore

Consideriamo di avere un insieme di N misure, $\{x_i\}_i^N$ di cui conosciamo il valore medio μ della popolazione e di volerne determinare la varianza, poichè essa dipende dalle misure del campione è una variabile aleatoria a sua volta e dunque da un campione definiamo una stima del valore reale del parametro σ_t. Di conseguenza possiamo domandarci le proprietà che tale stimatore possiede.
Definiamo lo stimatore varianza come:

$$\hat{\sigma}_\mu^2(x) = \frac{1}{N}\sum (x_i-\mu)^2 \quad \text{oppure} \quad \hat{\sigma}^2(x) = E(x^2) - E(x)^2$$

Verifichiamo che la varianza sia uno stimatore non distorto ovvero che:

$$E(\hat{\sigma}^2) = \sigma_t$$

Per farlo sfruttiamo la porprietà di linearità del valore atteso.
Dim.

$$E(\hat{\sigma}_\mu^2) = E(\frac{1}{N}\sum (x_i-\mu)^2) = \frac{1}{N}\sum(E(x_i^2) - 2\mu E(x_i) + \mu^2) =$$

$$= \frac{1}{N}\sum(E(x_i^2) - 2\mu E(x_i) + \mu^2) = \frac{1}{N}\sum(E(x_i^2) - \mu^2) = \frac{1}{N}\cdot N\cdot\sigma_t = \sigma_t$$

Dunque la varianza è uno stimatore non distorto nel caso in cui si conosca il valore medio della popolazione. Raramente si conosce μ della popolazione, dunque consideriamo come stimatore la varianza per un campione di N misure IID.

$$\hat{\sigma}_{\overline{x}}^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i-\overline{x})^2$$

Sappiamo che la varianza determina la dispersione di un campione di misure attorno alla sua media. Ipotizziamo di conoscere il valore medio del campione $\overline{x}$, per costruzione risulta essere il valore più vicino alle misure dell'insieme. La media della popolazione μ non necessariamente coincide con $\overline{x}$ del campione, e dunque può non essere il valore attorno al quale si distribuiscono le misure del campione; infatti nel caso in cui non lo sia al crescere del numero di misure questo può diventare il valore più distante rispetto a $\overline{x}$ stimato dal campione iniziale. Le distanze quadratiche da $\overline{x}$ saranno quindi una sottostima di μ e quindi anche $\hat{\sigma}$ sarà uno sottostima di σ_t.

$$E[\hat{\sigma}^2_{\overline{x}}] = \frac{1}{N}\sum E(x_i^2) - E([\frac{1}{N}\sum x_i]^2) = \sigma_t(x)^2 + \mu^2 - \frac{1}{N^2}[\sigma_t^2(\sum x_i) + E(\sum x_i)^2] =$$

$$= \sigma_t(x)^2 + \mu^2 - \frac{1}{N^2}[N\sigma_t^2(x) + N^2\mu^2] = \sigma_t^2(x)\Big[\frac{N-1}{N}\Big]$$

Di conseguenza la varianza di un campione $\hat{\sigma}_{\overline{x}}$ è uno stimatore distorto, infatti:

$$b_n[\hat{\sigma}^2_{\overline{x}}] = E[\hat{\sigma}^2_{\overline{x}}] - \sigma_t^2 = \sigma_t^2(x)\Big[\frac{N-1}{N}\Big] - \sigma_t^2$$

ma **asintoticamente non distorto** poichè per $N \to \infty$ si ha $b_n[\hat{\sigma}^2_{\overline{x}}] \to 0$.

Notare che quest'ultima definizione è quella operativa per verficiare che la varianza sia uno stimatore non ubiased in quanto difficilmente si conosce il valore medio μ della popolazione.

4.3.4 Correzione di Bessel

Si può definire un terzo stimatore, che introduca una correzione a $\hat{\sigma}^2_{\overline{x}}$ tale da cancellare il bias. La correzione del bias è applicabile tutte le colte in cui il bias è precisamente noto. Il nuovo stimatore della varianza sarà dato da:

$$s^2 \equiv \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \overline{x})^2 \tag{4.5}$$

e prende il nome di **correzione di Bessel**.
Lo stimatore è unbiased $E[s^2] = \sigma_t^2$, ma la varianza di tale stimare non può

essere determinata per il caso generale

$$V[s^2] = V[\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \overline{x})^2] = \frac{1}{(N-1)^2}\sum_{i=1}^{N}V[(x_i - \overline{x})^2]$$

ma è possibile farlo solo nel caso in cui il campione di misura segue una pdf(x) **Gaussiana**.

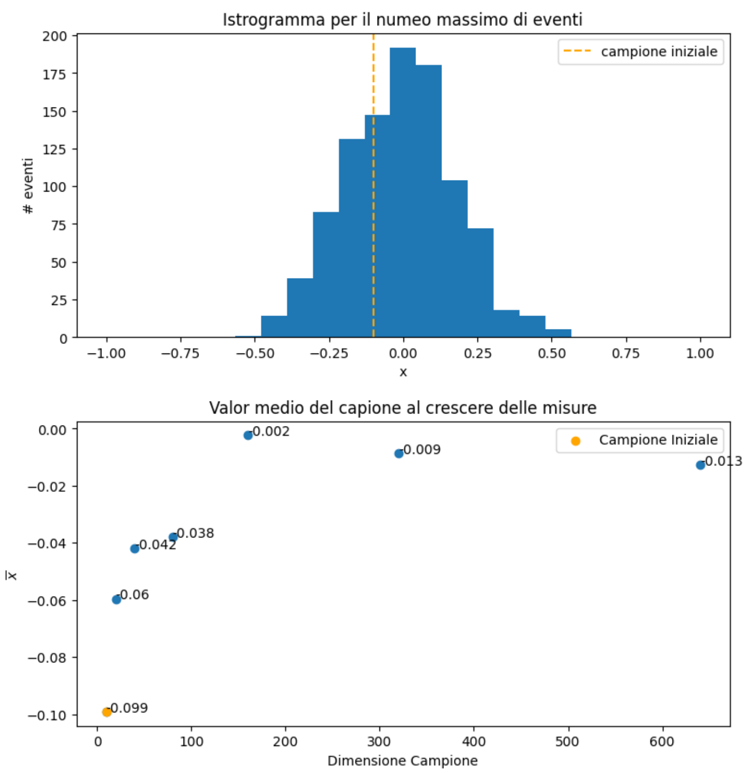


Figure 4.4: Misure casuali che seguono una pdf di Gauss tra -1 e 1 in cui $\overline{x}$ del campione iniziale non coincide con il valore medio $\mu = 0$ della popolazione. Dunque $\overline{x}$ non è più il centro del campione al crescere delle misure.

Se riscriviamo lo stimatore s^2 nel seguente modo:

$$s^2 \equiv \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \overline{x})^2 = \frac{\sigma_t^2}{N-1}\sum_{i=1}^{N}\frac{(x_i - \overline{x})^2}{\sigma_t^2}$$

possiamo introdurre una variabile aleatoria ausiliaria definita come χ^2 e riscrivere s^2 come:

$$s^2 \equiv \frac{\sigma_t^2}{N-1}\chi^2$$

di conseguenza $V[s^2]$ è legata alla $V[\chi^2]$. Nell'ipotesi in cui le misure raccolte seguano una distribuzione di probabilità Gaussiana la variabile χ^2 segue la distribuzione del chi-quadro. Tale pdf è descritta da un solo parametro definito gradi di libertà e nel nostro caso tale parametro vale N-1.
Di conseguenza con questa nuova distribuzione possiamo dimostrare che s^2 è uno stimatore non distorto.

Dim.

$$E[s^2] = E\left[\frac{\sigma_t^2}{N-1}\chi^2\right] = \frac{\sigma_t^2}{N-1}E[\chi^2] = \frac{\sigma_t^2}{N-1}(N-1) = \sigma_t^2$$

La sua varianza è data da:

$$V[s^2] = V\left[\frac{\sigma_t^4}{(N-1)^2}\chi^2\right] = \frac{\sigma_t^4}{(N-1)^2}V[\chi^2] = \frac{2\sigma_t^4}{(N-1)^2}$$

4.4 Intervallo di confidenza

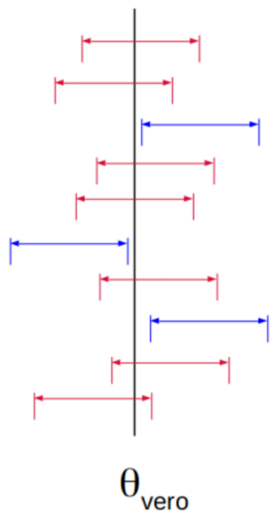
Si è costruito uno stimatore $\hat{\theta}$ per un parametro θ che è una variabile aleatoria di conseguenza esiste una distribuzione pdf($\hat{\theta}$) che è caratterizzata da una media E[$\hat{\theta}$] e una varianza V[$\hat{\theta}$]. La **stima** è il valore che lo stimatore assume in corrispondenza di un campione di dati e lo definiamo $\hat{\theta}^*$.

L'**intervallo di confidenza** è solitamente indicato come:

$$\hat{\theta}^* \pm \sigma \quad \text{oppure} \quad \hat{\theta}^{*}\,{}^{+\sigma_1}_{-\sigma_1} \tag{4.6}$$

la seconda notazione viene usata quando l'intervallo è asimmetrico.

In generale il risultato della stima è un intervallo [a,b], poichè è una variabile aleatoria si ha che gli estremanti sono variabili aleatorie e dunque per diversi campioni si hanno diversi intervalli di confidenza. A ciascun intervallo viene associata una probabilità (**livello di confidenza**) che misura quanto è buona la stima del valore vero.



$$P(\theta_t \in [a,b])$$

dove essa rappresenta la frazione delle volte in cui ripetendo l'esperimento, la stima restituisce un intervallo che contiene il valore vero.

In generale si vuole trovare un metodo che individui un intervallo [a,b] tale per cui la probabilità che $\theta_t \in [a,b]$ sia pari a un certo valore β detto **livello di confidenza**. Un intervallo di confidenza così definito prende il nome di **intervallo di condifenza di Neyman**. Un intervallo particolare è quello dato da $[\hat{\theta}^* - \sigma, \hat{\theta}^* + \sigma]$ e si definisce **intervallo centrale**.