Elementi di Matematica e di Statistica

Introduzione all'Inferenza Statistica

Docente: Riccardo Ievoli
riccardo.ievoli@unife.it

Corso di Laurea in Biotecnologie

19/11/2025

Indice

I metodi inferenziali si basano su tre concetti già introdotti:

- **Variabile** statistica
- **Popolazione** di riferimento
- **Campione** casuale

Si possono distinguere due aree all'interno dell'**inferenza statistica**:

1. **Stima**: l'obiettivo è conoscere caratteristiche della popolazione di riferimento . La stima si può suddividere in:
   - Stima puntuale (un solo valore)
   - Stima intervallare (un intervallo di valori)
2. **Test di ipotesi**: si ha un'ipotesi di lavoro sulla popolazione di riferimento che si vuole valutare in maniera rigorosa, partendo dai dati

Supponendo di osservare la variabile X, si definisce la sequenza di variabili che costituiscono il campione:

$$X_1, \ldots, X_i, \ldots, X_n.$$

Sono necessarie le seguenti **assunzioni**:

- Le unità statistiche sono **indipendenti**.
- La popolazione di riferimento è **infinita**.
- Ogni X_i è una variabile aleatoria: è quindi necessario specificare la sua legge di probabilità.

1. Ipotesi di lavoro: individuazione variabile/i e popolazione di interesse. Si fissa una numerosità n per il campione casuale:

$$X_1, \ldots, X_i, \ldots, X_n$$

2. Si assumono indipendenza, infinitezza di popolazione e un modello probabilistico parametrico:

$$Xi \sim f(x, \theta)$$

Le osservazioni sono *indipendenti* ed *identicamente distribuite* (i.i.d.).

Statistica Inferenziale

Gli "step" dell'Inferenza (2)

3 Si estrae il campione osservato

$$x_1, \dots, x_i, \dots, x_n.$$

4 Ci si interroga sulle **assunzioni**.

5 Si procede a fare inferenza: stima del parametro θ
A questo punto, si possono estendere i risultati ottenuti alla popolazione fornendo misure di **incertezza**.

- *Come si produce una stima del parametro di interesse θ?*

  Si impiega uno **Stimatore**.

- *Cos'è uno stimatore*?

  È una funzione dei dati campionari chiamata a volte anche *statistica*.

Lo stimatore è una **funzione** delle variabili aleatorie $X_1, \ldots, X_i, \ldots, X_n$

$$S = h(X_1, \ldots, X_i, \ldots, X_n).$$

Lo stimatore S è quindi una **variabile aleatoria**.

Il concetto di Stimatore (2)

Una volta osservate le caratteristiche sul campione osservato: $x_1, \ldots, x_i, \ldots, x_n$ si ottiene anche il valore campionario dello stimatore (o della statistica):

$$s = h(x_1, \ldots, x_i, \ldots, x_n).$$

Se questo stimatore è usato come stimatore del parametro θ, s sarà la stima di θ

Lo stimatore S è quindi una variabile aleatoria. Esso è un *buono* stimatore per θ se vengono rispettate alcune proprietà:

- Il valore atteso di S coincide con il valore vero (non noto) di θ. Quando questo accade si parla di stimatore **corretto** (o, meglio, **non distorto**).
$$E(S) = \theta$$
- La varianza di S è più bassa possibile. Quando uno stimatore ha una varianza inferiore ad un altro si dice **più efficiente**

Si distingue tra due tipologie di stime:

- **Stima puntuale**: valore singolo assunto dalla statistica campionaria (stimatore).
  Problema: fornire solo un valore puntuale implica ignorare che la stima risente di un certo livello di incertezza
- **Stima intervallare**: si propone l'uso di un intervallo di valori in cui ci si aspetta di trovare il vero parametro di popolazione, dato un certo *livello di confidenza*

L'assunzione più comune per la forma distributiva delle X_i è la distribuzione **Gaussiana** (o Normale):

$$X_i \sim N(\mu, \sigma^2)$$

Pertanto, i parametri che dobbiamo stimare sono **DUE**.

- La media μ
- La varianza σ^2

Stima Puntuale

Stima della media sotto l'assunzione di normalità (2)

Lo stimatore di μ con proprietà migliori è la *variabile aleatoria* media campionaria:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Se $X_1, \ldots, X_i, \ldots, X_n$ sono $N(\mu, \sigma^2)$, allora:

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

Osservazioni

- $\bar{X}$ è uno stimatore non distorto di μ poiché $E(\bar{X}) = \mu$ (dimostrazione nella prossima slide)
- La varianza dello stimatore cala all'aumentare di n.

Lo stimatore di μ con proprietà migliori è la *variabile aleatoria* media campionaria:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Se $X_1, \ldots, X_i, \ldots, X_n$ sono $N(\mu, \sigma^2)$, allora: $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$

Correttezza (Non distorsione)

$$E(\bar{X}) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n} \cdot E\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n} \cdot (n\mu) = \mu$$

Stima della media sotto l'assunzione di normalità (4)

Quindi la stima puntuale $\hat{\mu}$ di μ si ottiene dopo aver osservato il campione $x_1, \ldots, x_i, \ldots, x_n$:

$$\hat{\mu} = \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Osservazione: Perché si utilizza la media e non la mediana?

Bisogna inoltre stimare la varianza della distribuzione σ^2.
In questo caso, la forma funzionale dello stimatore è leggermente diversa rispetto a quanto visto in statistica descrittiva:

$$S^2 = \frac{1}{(n-1)} \sum_{i=1}^{n} (X_i - \bar{X})^2$$

Tale statistica (o stimatore) è chiamata **varianza campionaria corretta**, poiché $E(S^2) = \sigma^2$

Inoltre, si può dimostrare che tale stimatore si distribuisce come una variabile casuale continua chiamata Chi-quadrato, che si indica con χ^2.

Stima Puntuale

Stima della varianza sotto l'assunzione di normalità (2)

Quindi la stima puntuale ($s^2(x) = \hat{\sigma}^2$) di σ^2 si ottiene dopo aver osservato il campione $x_1, \ldots, x_i, \ldots, x_n$:

$$s^2(x) = \hat{\sigma}^2 = \frac{1}{(n-1)} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

Il punto di partenza per la costruzione di **intervalli di confidenza** per la media sotto normalità è il seguente risultato:

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim t_{(n-1)}$$

Ciò vale quando σ^2 (parametro di disturbo non noto) viene sostituito dalla sua stima $\hat{\sigma}^2$

Si noti come la variabile aleatoria nell'espressione sia $\bar{X}$, e μ sia il parametro di interesse che è quindi fisso ma <u>non noto</u>.

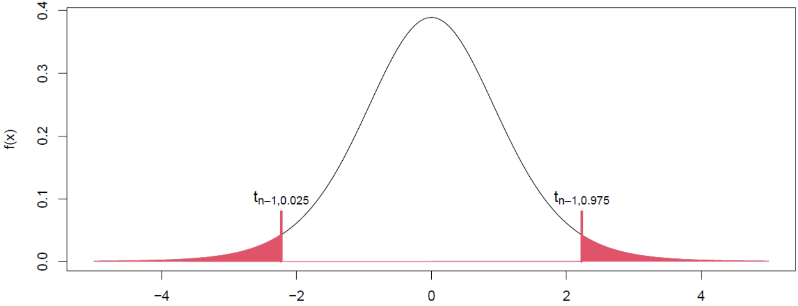
Stima Intervallare

Intervallo di confidenza della media sotto l'assunzione di normalità (2)

Dato un certo livello di confidenza $1-\alpha$, si dimostra che l'intervallo di confidenza per μ è:

$$\left[\bar{x} - \frac{\hat{\sigma}}{\sqrt{n}} t_{(n-1),(\alpha/2)}; \bar{x} + \frac{\hat{\sigma}}{\sqrt{n}} t_{(n-1),(1-\alpha/2)}\right]$$

Esempio con $1-\alpha = 0,95$



Stima Intervallare

Intervallo di confidenza della media sotto l'assunzione di normalità (3)

- Quando la numerosità campionaria è elevata, la t di Student è indistinguibile da una normale standard (approssimazione asintotica)
- In tal caso, l'intervallo di confidenza (approssimato) diventa il seguente (simmetria della Normale: $-z_{\alpha/2} = z_{1-\alpha/2}$):

$$\left[\bar{x} - \frac{\hat{\sigma}}{\sqrt{n}}|z_{\alpha/2}|; \bar{x} + \frac{\hat{\sigma}}{\sqrt{n}}|z_{1-\alpha/2}|\right]$$

- Inoltre, ci sono dei casi in cui σ^2 è noto e non da stimare. In quel caso si utilizzano sempre i quantili della normale (intervallo esatto).
- Nel caso in cui $1 - \alpha = 0,95$, allora $z_{\alpha/2} = -1,96$ e $z_{1-\alpha/2} = +1,96$.
- **Regola empirica** per costruire l'intervallo (approssimato) al 95%:

$$\bar{x} \pm 2 \cdot \frac{\hat{\sigma}}{\sqrt{n}}$$

Cosa abbiamo imparato oggi?

1. Stima puntuale
2. Stima intervallare

Si registrano i livelli di *acido palmitico* (C16:0) su 10 tipologie di biscotti ottenendo la seguente distribuzione:

$$\{16,1; 20,1; 18,1; 9,5; 10,0; 9,9; 43,5; 38,8; 41,7; 28,5\}$$

Assumendo che la distribuzione sia gaussiana:

1. si calcoli la stima della media $\bar{x}$
2. si calcoli la stima della varianza $\hat{\sigma}^2$
3. Qual è l'intervallo di confidenza per la media al 95% ipotizzando che $\sigma^2 = \hat{\sigma}^2$?

Stima Puntuale ed Intervallare

Esercizi (2)

ESERCIZIO 1. Un ricercatore è interessato a stimare il livello medio di un dato enzima in una certa popolazione umana e dispone di un campione di $n = 10$ individui che presentano una media pari a $\bar{x} = 10$. Supponendo che la variabile di interesse si distribuisca come una normale con varianza pari a $\sigma^2 = 25$, si determini l'intervallo di confidenza al 95% per la media.

ESERCIZIO 2. Un fisioterapista vuole stimare, con un livello di fiducia del 95%, la forza media massima di un particolare muscolo in un certo gruppo di individui. Egli assume che i punteggi di forza si distribuiscano come una normale con varianza pari a 144. L'estrazione di un campione casuale di 15 soggetti ha portato ad un valore medio del punteggio pari a 84,3.

ESERCIZIO 2.1. Si stimi l'intervallo di confidenza per l'esempio precedente ma con un livello di fiducia del 99%

Stima Puntuale ed Intervallare

Esercizi (3)

ESERCIZIO 3 con Excel. Considerando il dataset "Pazienti" (n=50), calcolare media ($\overline{x}$) e varianza campionarie ($s^2(x)$ o $\hat{\sigma}^2$) (occhio alla divisione per n-1 e non per n) nonché l'intervallo di confidenza al 95% (considerare $|z_{\alpha/2}| = |z_{0,025}| = 1,96$) per la media delle seguenti variabili:

- Età
- FEV1
- DLCO
- Hospital Stay (giorni)
- Operation Time (minuti)