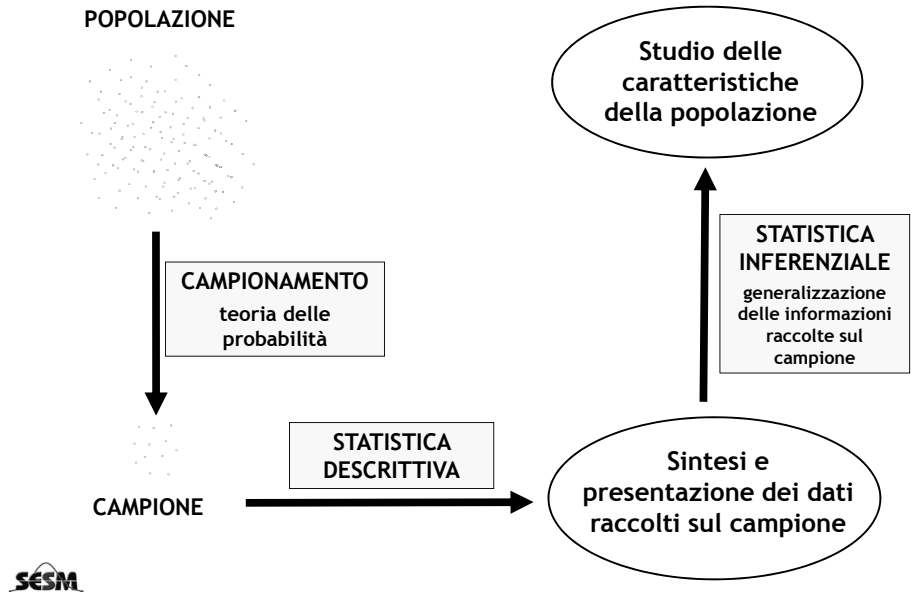
STATISTICA INFERENZIALE

SCHEMA LOGICO DELLA STATISTICA

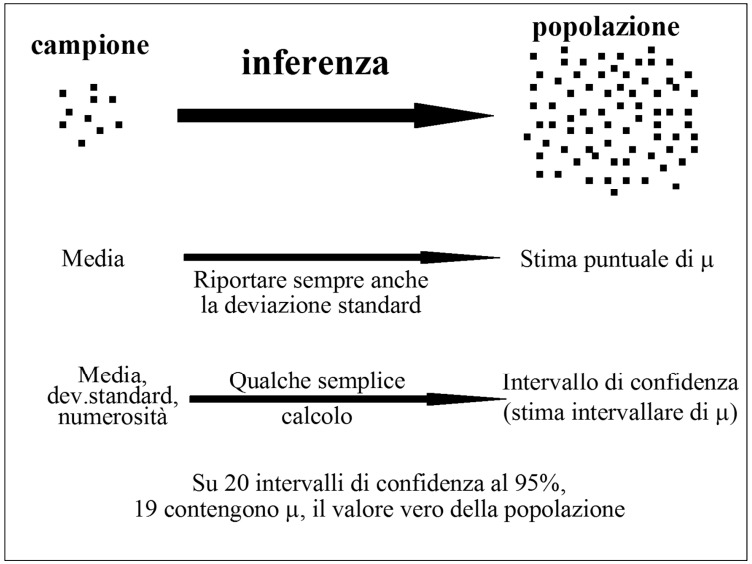


Statistica Inferenziale

Le domande a cui la Statistica Inferenziale cerca di dare una risposta rientrano in 3 principali categorie

1. Si vuole stimare il vero valore dei parametri ignoti in base alle osservazioni del campione e capire quanto accurata è la stima proposta (STIMA PUNTUALE).
2. Si vuole identificare un insieme di valori ragionevoli per i parametri ignoti (STIMA INTERVALLARE).
3. Si formula un'ipotesi sul vero valore dei parametri ignoti e si vuole verificare se tale ipotesi è vera oppure no, in base alle osservazioni capionarie. (VERIFICA DI IPOTESI).

Brevi cenni all'intervallo di confidenza



Dal momento che il campione viene estratto casualmente dalla popolazione, le conclusioni tratte da un campione possono essere errate.

L'inferenza statistica viene fatta "con umiltà":

1) si cerca di stimare la probabilità di commettere errori

2) si cerca di limitare la probabilità di commettere errori

INTERVALLO di CONFIDENZA

Lo scopo dell'inferenza statistica è la conoscenza dei **parametri** che caratterizzano una popolazione.

Per conoscere il parametro, però, dovremmo prendere in esame **tutte** le unità statistiche che costituiscono la popolazione; questo spesso è impossibile perché:

1. numerosità molto elevata
2. spesso la popolazione obiettivo è infinita

impossibile conoscere il **parametro**

Non potendo calcolare con esattezza il parametro, **ricorriamo ad una sua stima**.

La **statistica** (es $\bar{x}$, s) calcolata su un campione estratto dalla popolazione obiettivo è una **stima puntuale** del parametro della popolazione.

Questa stima puntuale del parametro non sarà mai identica al vero parametro della popolazione, ma sarà affetta da un **errore** per eccesso o per difetto.

In molte situazioni è preferibile **una stima intervallare** (cioè è preferibile indicare come stima del parametro un intervallo al posto di un *singolo punto* sull'asse dei valori) che esprima anche l'**errore associato alla stima** (precisione).

INTERVALLO di CONFIDENZA: DEFINIZIONE

Per intervallo di confidenza di un parametro Θ (ad es. della media μ) della popolazione, intendiamo un intervallo delimitato da due limiti L_{inf} (limite inferiore) ed L_{sup} (limite superiore) che abbia una definita probabilità $(1-\alpha)$ (ad es. (1-0.05)=0.95) di contenere il vero parametro della popolazione:

$$p(L_{inf} < \Theta < L_{sup}) = 1-\alpha$$

$$p(L_{inf} < \mu < L_{sup}) = 1-0.05 = 0.95$$

dove:

$1-\alpha$ = grado di confidenza

α = probabilità di errore

Esempio

- Vogliamo stimare il livello medio di glicemia nei diabetici italiani:
- prendiamo un campione di 36 soggetti; la media della glicemia in questo gruppo risulta 155 mg/dl *(stima puntuale).*
- Calcoliamo l'intervallo di confidenza al 95% che risulta: 147,2-162,8 mg/dl *(stima intervallare)*

questo intervallo ha una probabilità del 95% di contenere la vera media della popolazione dei diabetici italiani

Esempio - *continua*

- L'intervallo di confidenza al 95% del lucido precedente è stato ottenuto nel seguente modo:

$$\bar{x} = 155 mg/dl$$

$$s = 24 mg/dl$$

$$n = 36$$

$$\bar{x} \pm 1{,}96 * \frac{s}{\sqrt{n}} = 155 \pm 1{,}96 * \frac{24}{\sqrt{36}}$$

L_{inf}=147,2

L_{sup}=162,8

RIASSUMENDO…

La **stima puntuale** fornisce un singolo valore. Tuttavia:

1. questo valore non coincide quasi mai con il valore vero (parametro) della popolazione;
2. campioni diversi forniscono stime puntuali diverse.

La **stima intervallare** fornisce un intervallo:

1. quest'intervallo ha una determinata probabilità (in genere, il 95%) di contenere il valore vero (parametro) della popolazione;
2. Il metodo generale per la costruzione dell'intervallo di confidenza al (1-α) è:

$$\bar{x} \pm z_{\alpha/2} \cdot \frac{s}{\sqrt{n}}$$

$$\bar{x} \pm z_{\alpha/2} \cdot \frac{s}{\sqrt{n}}$$

la **probabilità d'errore** α determina il valore del coefficiente z:

1-α	$\alpha/2$	$z_{\alpha/2}$
0.90	0.05	1.64
0.95	0.025	1.96
0.99	0.005	2.58

Cenni all'interpretazione di un Test d'ipotesi

TEST D'IPOTESI

⇒ In medicina una delle più utilizzate tecniche inferenziali è quella nota come *test d'ipotesi.*

Tale procedura è particolarmente utile in situazioni in cui noi siamo interessati a *prendere decisioni tra due o più alternative possibili*, piuttosto che alla stima del valore di uno o più parametri.

Ad esempio

- valutare l'efficacia di un nuovo farmaco rispetto al placebo
- valutare se il trattamento chirurgico di un particolare tumore in una data fase allunga la vita dei pazienti rispetto al trattamento chemioterapico
- valutare se l'esposizione a una determinata sostanza chimica è responsabile di un eccesso di tumori

In tali situazioni la valutazione dell'alternativa migliore è finalizzata a decidere quale intervento operare sulla realtà
(scelta del farmaco, tipo di terapia, tipo di intervento preventivo)

Es.: Ci si chiede se la glicemia dei diabetici italiani sia diversa dalla glicemia dei diabetici americani.

Nei diabetici americani il livello medio di glicemia della popolazione è pari a 170 mg/dl (e la deviazione standard è 24 mg/dl).

In un gruppo di 36 diabetici italiani il livello medio di glicemia è di 162 mg/dl.

La differenza che osserviamo tra i due valori di glicemia è solo dovuta al caso o rispecchia una differenza presente in realtà tra le due popolazioni?

TEST D'IPOTESI

Tutte le differenze osservate sono delle semplici fluttuazioni casuali

Le differenze riscontrate nei campioni rispecchiano una reale differenza nelle popolazioni corrispondenti

Esempio:

La glicemia dei diabetici italiani è uguale alla glicemia dei diabetici americani

La glicemia dei diabetici italiani è diversa dalla glicemia dei diabetici americani

Dati del campione

Test statistico

P>0,05 = la probabilità che le differenze osservate siano dovute al caso è superiore al 5%

P<0,05 = la probabilità che le differenze osservate siano dovute al caso è inferiore al 5%

le differenze osservate tra i campioni possono essere attribuite al caso

le differenze osservate tra i campioni rispecchiano delle differenze reali tra le popolazioni

Si effettua il test statistico:

TEST STATISTICO → $z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$

$$z = \frac{162 - 170}{24 / \sqrt{36}} = -2$$

Consultando le tabelle della normale standardizzata, si trova il valore di P (per il valore ottenuto di -2):

P=0,046 → P <0,05

Quindi la media della glicemia nella popolazione italiana è diversa da quella della pop. americana

- quando la probabilità P< 5%, si dice che c'è una differenza statisticamente significativa (ad es. tra la glicemia dei diabetici americani e quella dei diabetici italiani)
- spesso si possono trovare anche le notazioni:
  - P<0,01: la probabilità che la differenza sia dovuta al caso è inferiore all' 1%
  - P<0,001: la probabilità che la differenza sia dovuta al caso è inferiore all' uno per mille
  - n.s. : differenza non statisticamente significativa; la probabilità che la differenza sia dovuta al caso è maggiore del 5%