Inferenza statistica

	Task Description	Due Date	Status
☐	Introduzione all'inferenza statistica: popolazione e campione, obiettivi, cenni ai metodi di campionamento		
☐	Cenni ai principali schemi di campionamento		
☐	Il campionamento casuale semplice: schema con e senza ripetizione		
☐	Campione casuale e campione osservato. Le v.c. osservazioni campionarie		
☐	Statistica: definizione, principio di naturalità, sufficienza		
☐	Gli spazi dell'inferenza statistica: spazio parametrico, spazio campionario e spazio della statistica		
☐	Distribuzione campionaria di una statistica		
☐	Distribuzione campionaria della media campionaria		
☐	Distribuzione campionaria della proporzione campionaria		
☐	Distribuzione campionaria di una statistica: un semplice esempio		

Parte descrittiva → insieme dati POPOLAZIONE

↓

obiettivo sintesi

Parte inferenziale → sottoinsieme della POPOLAZIONE

↓

campione

↓

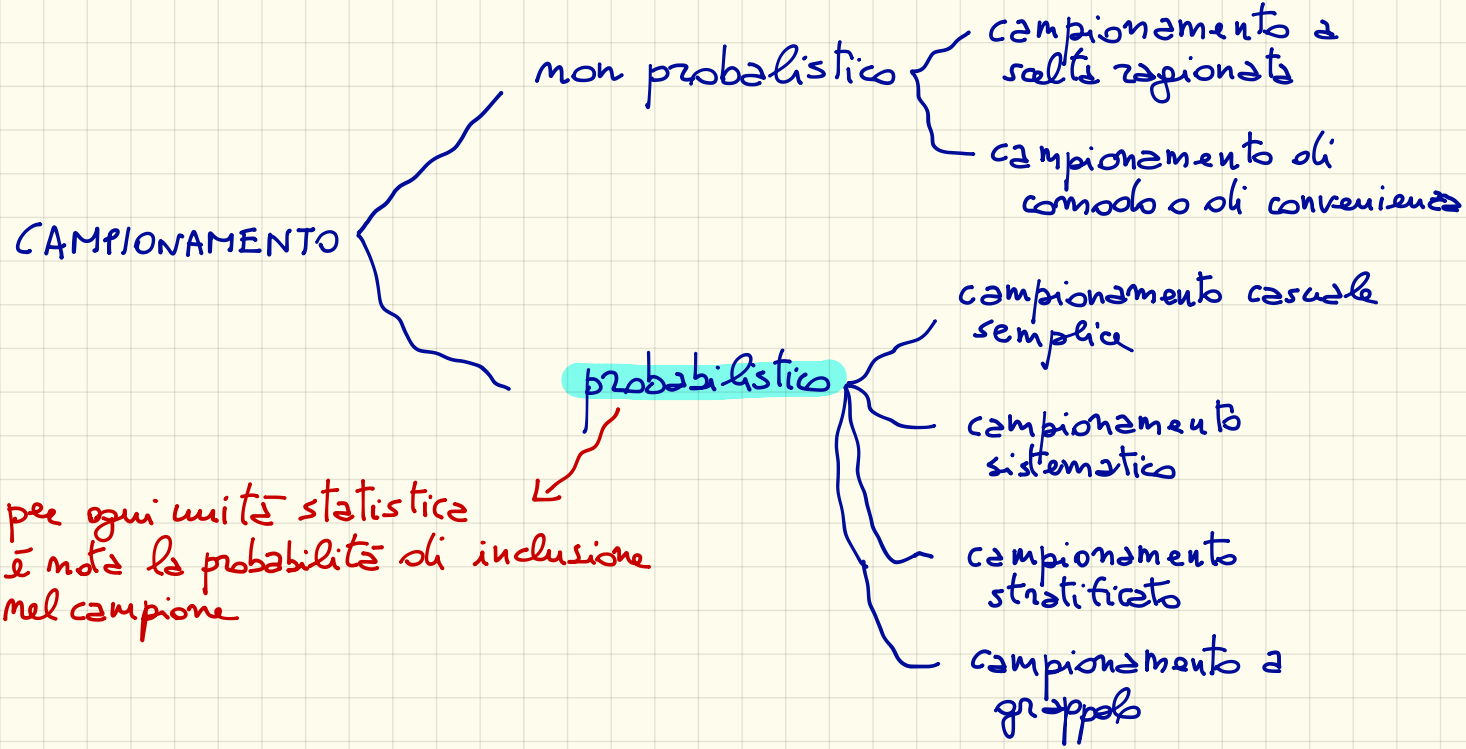
sintesi del campione

obiettivo diretto ad ottenere informazioni sulla popolazione

se il campione è estratto casualmente gli strumenti della probabilità ci aiutano in questo passaggio inverso

↓

INFERENZA


CAMPIONAMENTO
non probabilistico
campionamento a scelta ragionata
campionamento di comodo o di convenienza
probabilistico
campionamento casuale semplice
campionamento sistematico
campionamento stratificato
campionamento a grappolo
per ogni unità statistica è nota la probabilità di inclusione nel campione

campionamento casuale semplice: tutte le unità statistiche della popolazione hanno la stessa probabilità di inclusione nel campione

- con ripetizione (più "naturale")
- senza ripetizione (vantaggio: semplicità)

se $N \to \infty$ i due schemi sono equivalenti

le unità della popolazione possono essere assimilate alle palline contenute in un'urna

→ [da un'urna che contiene N palline ne estraggo casualmente n] { con ripetizione; senza ripetizione }

campionamento stratificato ⇝ la popolazione è caratterizzata da un certo numero di strati

riconducibili alle differenti modalità di una variabile qualitativa

es: genere ⇝ M, F

es: diploma ⇝ LC, LS, IT, IP, Altro

3 livelli o strati corrispondono a varie differenti

↓

il processo di campionamento cercherà di riprodurre nel campione le proporz. degli strati presenti nella popolazione

MEMO COMBINAZIONE LINEARE V.C.

Sia $X_1, X_2, \dots, X_n$ una n-pla di r.c. tali che le X_i siano i.i.d. come una $X \sim f(x;\theta)$.

Sia inoltre $\mu_X = E(X) = E(X_i)$ e $\sigma^2 = Var(X) = Var(X_i)$.

Sia W una combinazione lineare delle n r.c. con pesi a_i:

$$W = \sum_i a_i X_i$$

Si ha:

$$\mu_W = E(W) = \mu \sum_i a_i \qquad \sigma_W^2 = Var(W) = \sigma^2 \sum_i a_i^2$$

Due particolari combinazioni lineari:

- somma $(a_i = 1 \ \forall i) \rightsquigarrow W = \sum_i X_i \begin{cases} \mu_W = n\mu \\ \sigma_W^2 = n\sigma^2 \end{cases}$
- media $(a_i = \frac{1}{n} \ \forall i) \rightsquigarrow W = \frac{\sum_i X_i}{n} = \bar{X}_n \begin{cases} \mu_W = \mu \\ \sigma_W^2 = \frac{\sigma^2}{n} \end{cases}$

Cosa è possibile dire circa la distribuzione di W?

Se $X \sim Ber(\pi)$ $\rightsquigarrow$ $W = \sum X_i \sim bin(n, \pi)$ $\quad E(W) = n\pi$, $Var(W) = n\pi(1-\pi)$

$\rightsquigarrow$ $W = \frac{\sum X_i}{n}$ trasformazione di una v.c. binomiale $\quad E(W) = \pi$, $Var(W) = \frac{\pi(1-\pi)}{n}$

Se $X \sim N(\mu, \sigma^2)$ $\rightsquigarrow$ $W = \sum a_i X_i \sim N(\mu_x \sum_i a_i, \sigma_x^2 \sum_i a_i^2)$

proprietà riproduttiva v.c. normale (distribuzione esatta)

Se $X \sim f(x;\theta)$ $\rightsquigarrow$ $W = \sum_i a_i X_i \xrightarrow[n\to\infty]{} N(\mu_x \sum_i a_i, \sigma_x^2 \sum_i a_i^2)$

f potrebbe non essere normale oppure non si hanno le informazioni necessarie per assumere la normalità

teorema del limite centrale (distribuzione approssimata)

POPOLAZIONE → si è interessati a studiare un fenomeno che si assume possa essere descritto da un modello probabilistico

ATTENZIONE: tutta l'inferenza si basa su queste ipotesi → $X \sim f(x; \theta)$

θ → parametro di interesse

↓

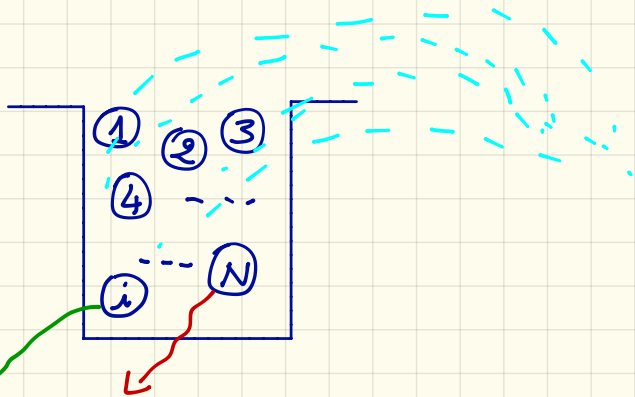
l'inferenza parametrica mira ad ottenere informazioni su θ a partire dai dati campionari

H

Spazio parametrico

↓

insieme dei possibili valori che può assumere il parametro θ



N: numerosità della popolazione

($N \to \infty$: popolazioni infinite)

[l'i-esima unità della popolazione]

CAMPIONE OSSERVATO

$x_1, x_2, \dots, x_i, \dots, x_m$

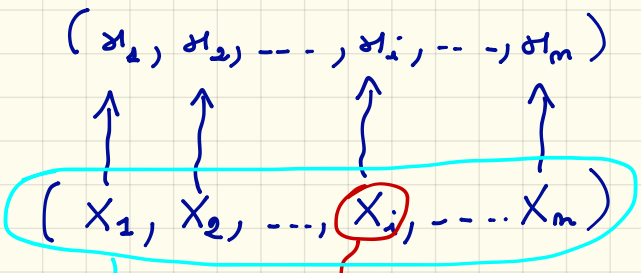
m: numerosità del campione

valore che la variabile X assume sull'i-esima unità del campione

il valore x_i osservato è uno dei possibili valori che possono essere osservati all'i-esima posizione del campione

x_i può essere considerato uno dei valori che una v.c. X_i può assumere

all'i-esima posizione del campione potrei osservare una qualunque delle unità della popolazione



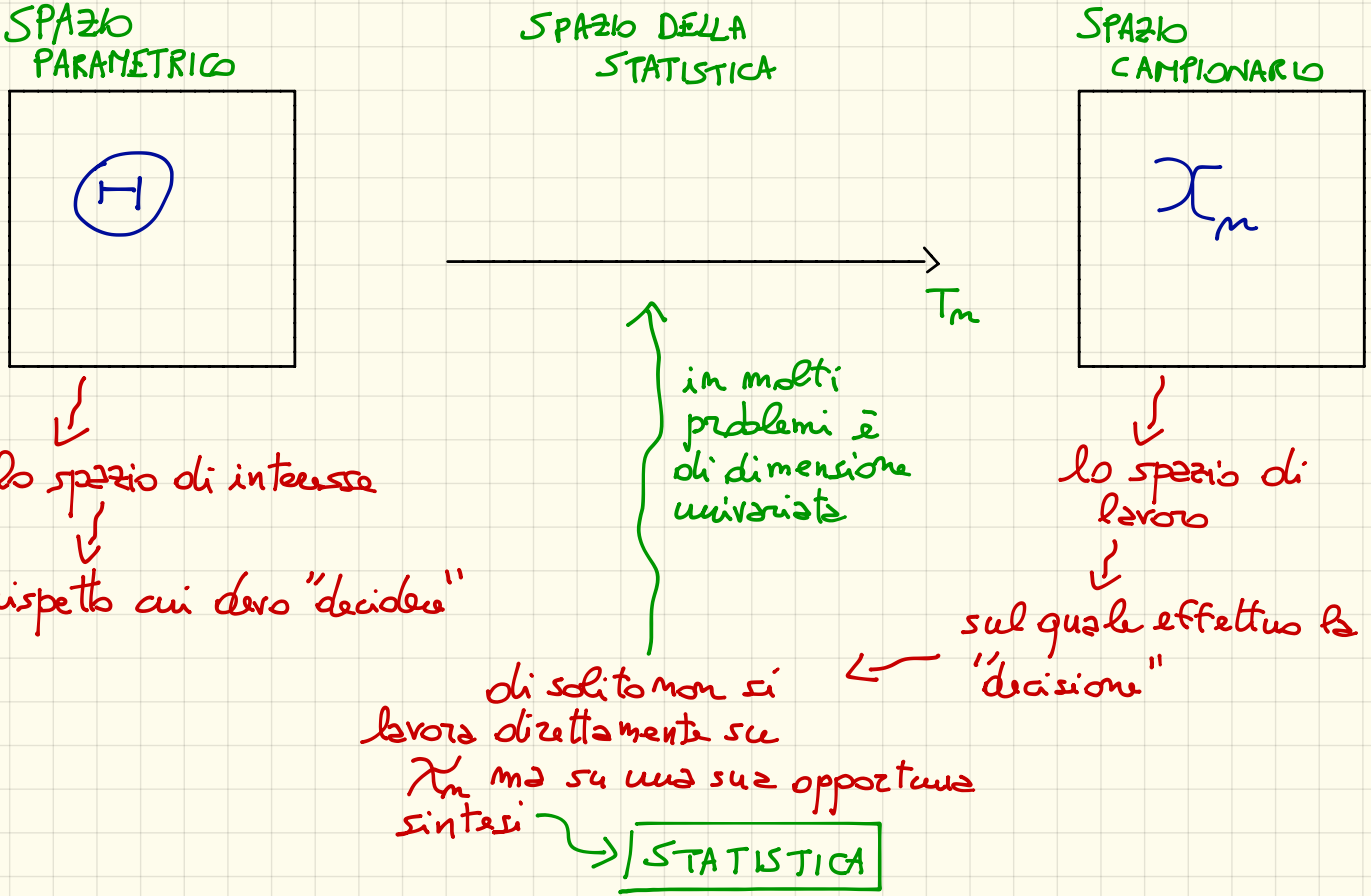
ciascuna $X_i \sim X \sim f(x;\theta)$

inoltre se il campionamento è con ripetizione le X_i sono indipendenti → le X_i sono i.i.d. $\sim X$

parallelo con le combinazioni lineari di v.c. richiamato ad inizio lezione

l'insieme di tutti i possibili campioni di ampiezza n (ovvero l'insieme di tutti i possibili valori che possono comporre il campione osservato) si chiama spazio campionario e si indica con $\mathcal{X}_n$ → NOTA: da non confondere con Ω (spazio campione)

GLI "SPAZI" DELL'INFERENZA



GLI "SPAZI" DELL'INFERENZA

la distribuzione di probabilità di T_n si chiama distribuzione campionaria

qualunque funzione del campione casuale che non dipende da parametri incogniti

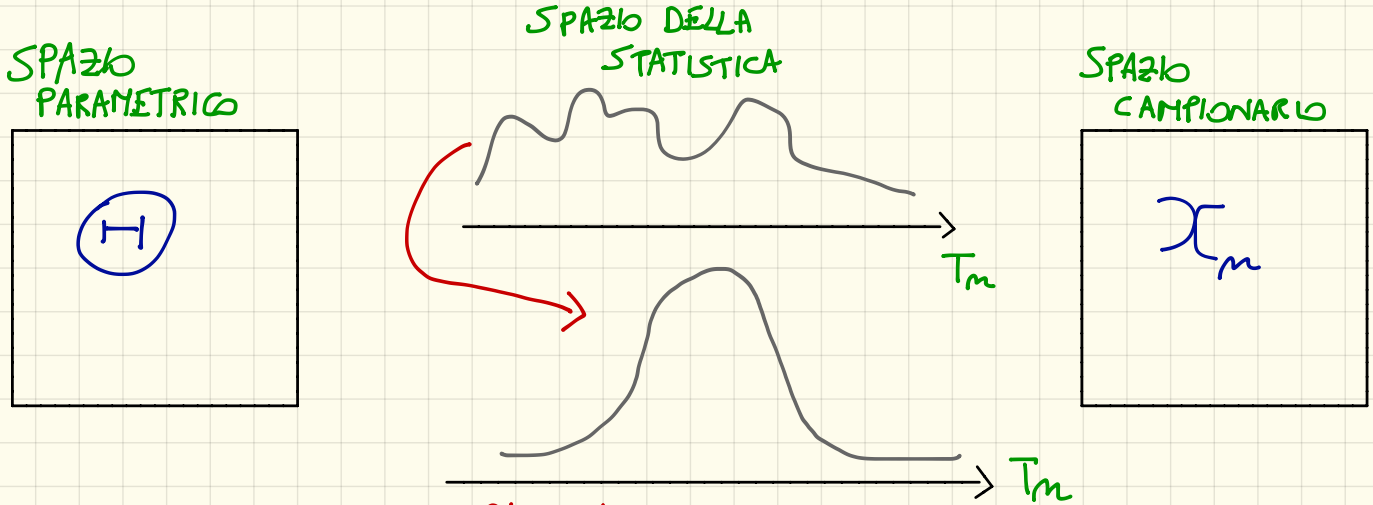
e che quindi può essere calcolata

$$T_n = T(X_1, X_2, \dots, X_i, \dots, X_n)$$

è una v.c. (funzione di n v.c.)

$$t_n = T(x_1, x_2, \dots, x_i, \dots, x_n)$$

valore assunto dalla statistica in corrispondenza del campione osservato



in molti problemi di interesse è possibile fare riferimento a statistiche caratterizzate da distribuzioni campionarie normali

$X \sim Bee(\pi)$

$T_n = \hat{\pi} = \frac{\sum X_i}{n} \rightarrow N\left(\pi, \frac{\pi(1-\pi)}{n}\right)$

distribuzione approssimata (TLC)

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

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

distrib. esatta (proprietà riproduttiva normale)

$X \sim f(x; \theta)$

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

distribuzione approssimata (TLC)

SULLA SCELTA DELLA "MIGLIORE" STATISTICA T_n

esistono una serie di proprietà formali che consentono di scegliere tra statistiche alternative

(tali proprietà dipendono anche dal particolare utilizzo che si deve fare di T_n

- problemi di stima → T_n si chiama stimatore
- verifica di ipotesi → T_n si chiama statistica test

due requisiti essenziali:

- naturalità — applico la stessa statistica che adopererei sulla popolazione all'unico campione disponibile

- SUFFICIENZA → il requisito essenziale è che T_n sia in grado di catturare tutte le informazioni utili su θ contenute nel campione

una volta calcolata la statistica T_n non è necessario conservare il campione

Inferenza su μ [DISTRIBUZIONE MEDIA CAMPIONARIA]

$T_n = \bar{X}_n = \frac{\sum X_i}{n}$ si usa la statistica media campionaria

- se $X \sim N(\mu, \sigma^2) \Rightarrow \bar{X}_n \sim N\left(\mu, \frac{\sigma^2}{n}\right)$ proprietà riproduttiva v.c. normale
- se $X \sim f(x;\theta) \Rightarrow \bar{X}_n \to N\left(\mu, \frac{\sigma^2}{n}\right)$ teorema del limite centrale

X non è normale oppure non è possibile assumere la normalità

$n \to \infty$

↳ REGOLA EMPIRICA: se la dimensione del campione (n) è ≥ 30 posso sfruttare il TLC

Inferenza su π [DISTRIBUZIONE PROPORZIONE CAMPIONARIA]

$T_n = \mathcal{F}_n = \frac{\sum X_i}{n}$ si usa la statistica proporzione campionaria

$\left(\overline{X}_n \right.$ si tratta di una media di osservazioni $\{0, 1\}$ $\left.\right)$

$\{0, 1\}$ → possesso della caratteristica di interesse (successo)

In questo caso: $X \sim Ber(\pi)$ → proporzione nella popolazione

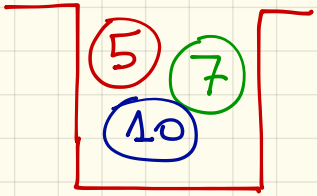
$\mathcal{F}_n = \hat{\pi}$ è una trasformazione di una v.c. binomiale

→ $\mathcal{F}_n = \frac{\sum X_i}{n} \rightarrow bin(n, \pi)$] è possibile fare riferimento al modello binomiale per ottenere la distribuzione esatta di $\mathcal{F}_n$

Essendo $\mathcal{F}_n$ una combinazione lineare di v.c. i.i.d. $\sim Ber(\pi)$ è possibile sfruttare il teorema del limite centrale

$\hat{\pi} \xrightarrow[n \to \infty]{} N\left(\pi, \frac{\pi(1-\pi)}{n}\right)$ DISTRIBUZIONE ASINTOTICA DI $\mathcal{F}_n$

CASO STUDIO (ci permette di introdurre i concetti basilari)



popolazione → N = 3

X n. errori per pagina di tre dattilografi

X	$p(x)$
5	1/3
7	1/3
10	1/3
	1

modello che descrive la popolazione

$E(X) = 7,33$

$Var(X) = 4,22$

due "parametri" caratteristici della popolazione

Ipotizziamo di estrarre un campione di numerosità

$n = 2$

(?) quanti sono i possibili campioni che potrei osservare (?)

$3 \times 2 = 6$

estrazione senza rimessa

$3 \times 3 = 9$

estraz. con rimessa

$X_1 \rightsquigarrow$ risultati che potrei osservare alla prima estrazione

$X_2 \rightsquigarrow$ risultati che potrei osservare alla seconda estrazione

X_1	X_2
5	7
5	10
7	5
7	10
10	5
10	7
5	5
7	7
10	10

[questi tre campioni si hanno nel caso di estrazione con rimessa

$X_1 (= X_2)$	senza rimessa	con rimessa
5	$2/6 = 1/3$	$3/9 = 1/3$
7	$2/6 = 1/3$	$3/9 = 1/3$
10	$2/6 = 1/3$	$3/9 = 1/3$

le singole X_i (osservazione campionaria) si distribuisce allo stesso modo della X che definisce la popolazione

→ X_i è i.d. come la X

nel caso di estrazioni con rimessa le X_i saranno anche indipendenti: X_i è i.i.d. come la X

X_1	X_2
5	7
5	10
7	5
7	10
10	5
10	7
5	5
7	7
10	10

parametro o vettore di parametri che caratterizzano f

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

fenomeno di interesse popolazione

nell'es. è il n. di errori commessi dalle dattilografe

modello di probabilità che si pensa possa descrivere adeguatamente il fenomeno

tutto quello che vedremo d'ora in poi si basa sull'ipotesi che la f sia adeguata

X_1	X_2	$\bar{X}_m$
5	7	6
5	10	7,5
7	5	6
7	10	8,5
10	5	7,5
10	7	8,5
5	5	5
7	7	7
10	10	10

principio di naturalità

↓

applico la stessa statistica che adopererei sulla popolazione all'unico campione disponibile

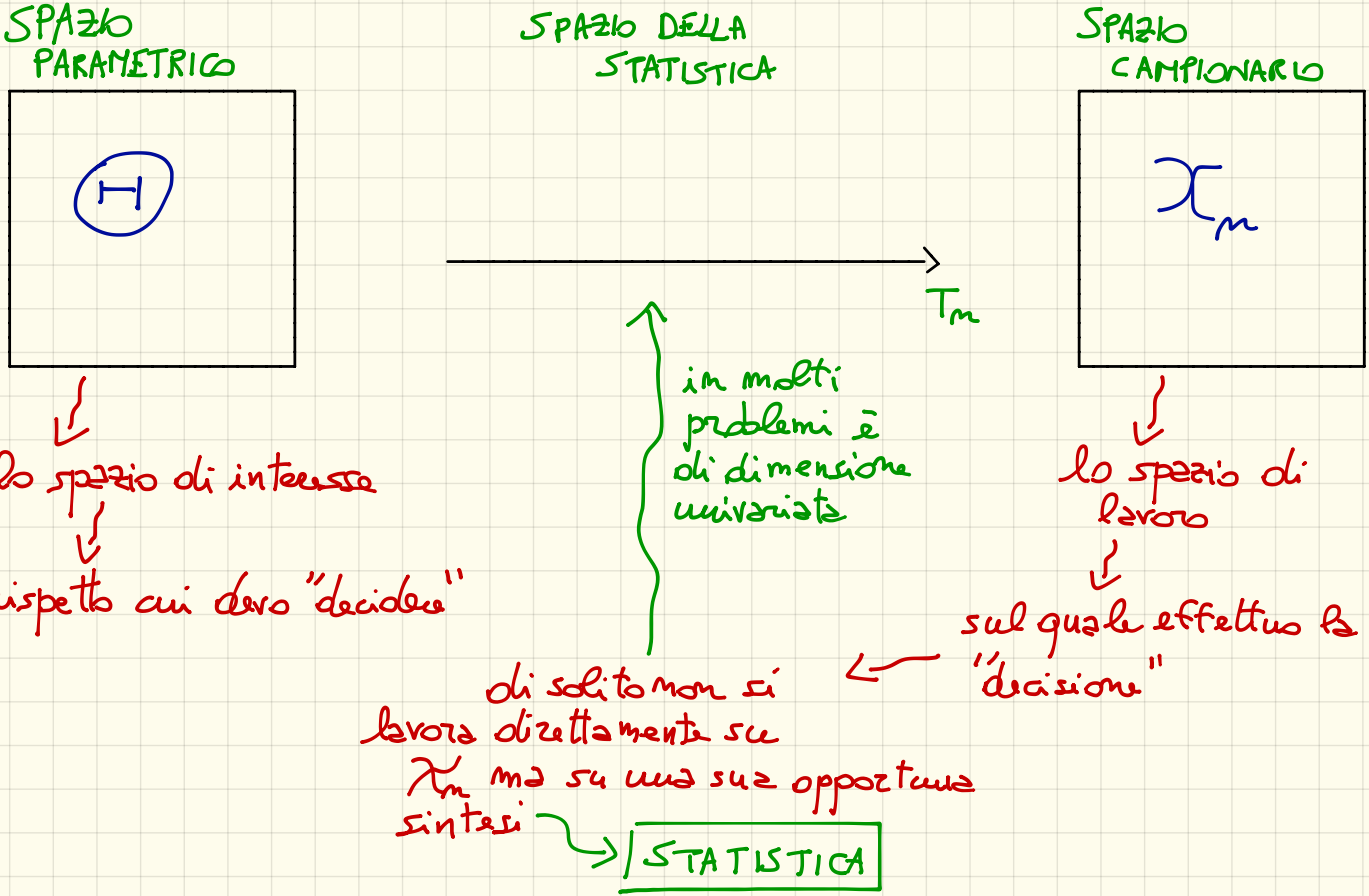
→ distribuzione campionaria della statistica: comportamento di $\bar{X}_m$ su tutti i possibili campioni di numerosità $n = 2$

Immaginiamo di essere interessati a studiare la media della popolazione

	Task Description	Due Date	Status
☐	I problemi dell'inferenza statistica: stima e verifica di ipotesi		
☐	Proprietà degli stimatori: la non distorsione.		
☐	Lo stimatore s^2 per la varianza Efficienza relativa ed efficienza assoluta		
☐	Criterio dell'errore quadratico medio come misura congiunta della posizione e della variabilità della distribuzione campionaria		
☐	Proprietà finite ed asintotiche di uno stimatore		
☐	Consistenza di uno stimatore		
☐	RIEPILOGO DISTRIBUZIONI CAMPIONARIE CASO DI INFERENZA SU UNA SOLA POPOLAZIONE		
☐	Inferenza sulla media: caso di varianza nota, di varianza incognita e rimozione dell'ipotesi di normalità		
☐	Inferenza sulla varianza di una popolazione normale		
☐	Inferenza sulla proporzione di una popolazione bernoulliana		

GLI "SPAZI" DELL'INFERENZA

MEMO



GLI "SPAZI" DELL'INFERENZA

MEMO

la distribuzione di probabilità di T_n si chiama distribuzione campionaria

qualunque funzione del campione casuale che non dipende da parametri incogniti

e che quindi può essere calcolata

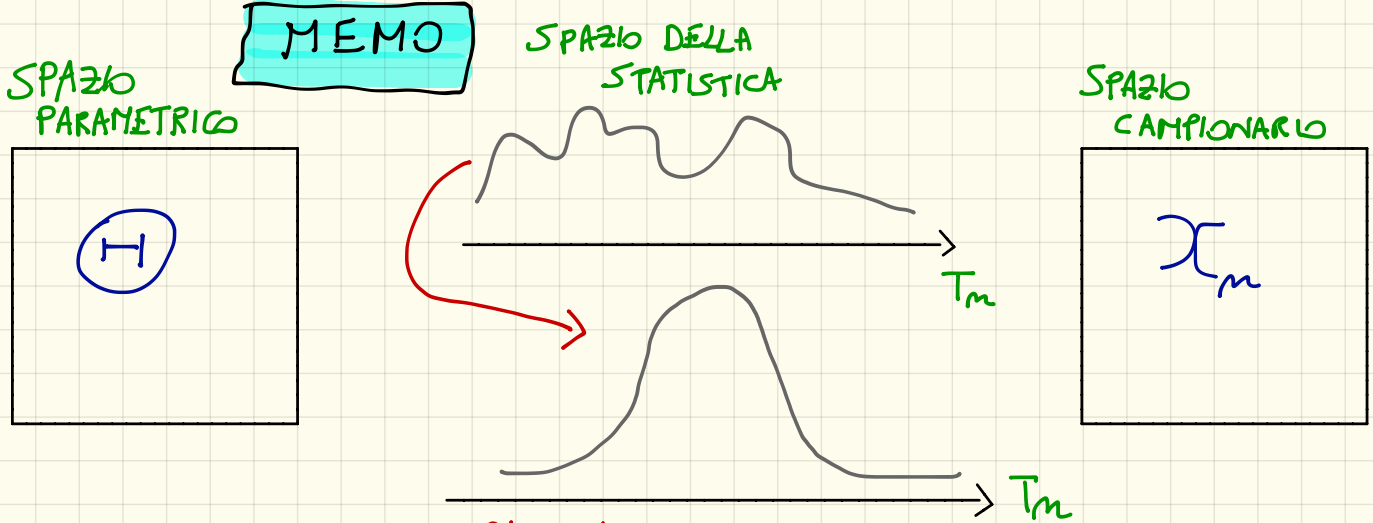
$$T_n = T(X_1, X_2, \dots, X_i, \dots, X_n)$$

è una v.c. (funzione di n v.c.)

$$t_n = T(x_1, x_2, \dots, x_i, \dots, x_n)$$

valore assunto dalla statistica in corrispondenza del campione osservato

MEMO



in molti problemi di interesse è possibile fare riferimento a statistiche caratterizzate da distribuzioni campionarie normali

$X \sim Bee(\pi)$

$T_n = \hat{\pi} = \frac{\sum X_i}{n} \to N\left(\pi, \frac{\pi(1-\pi)}{n}\right)$

distribuzione approssimata (TLC)

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

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

distrib. esatta (proprietà riproduttiva normale)

$X \sim f(x; \theta)$

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

distribuzione approssimata (TLC)

SULLA SCELTA DELLA "MIGLIORE" STATISTICA T_n

MEMO

esistono una serie di proprietà formali che consentono di scegliere tra statistiche alternative

(tali proprietà dipendono anche dal particolare utilizzo che si deve fare di T_n)

- problemi di stima → T_n si chiama stimatore
- verifica di ipotesi → T_n si chiama statistica test

due requisiti essenziali:

- naturalità — applico la stessa statistica che adopererei sulla popolazione all'unico campione disponibile

- SUFFICIENZA → il requisito essenziale è che T_n sia in grado di catturare tutte le informazioni utili su θ contenute nel campione

una volta calcolata la statistica T_n non è necessario conservare il campione

INFERENZA (parametrica)

problemi di stima

si vuole stimare il parametro θ a partire da osservazioni campionarie

la statistica T_n, sintesi delle osservazioni campionarie, si chiama **STIMATORE**

verifica di ipotesi

si vuole verificare la validità di un'ipotesi su θ a partire dall'evidenza campionaria

la statistica T_n si chiama in questo caso **STATISTICA TEST**

in entrambi i casi la scelta della migliore statistica T_n da usare si basa su una serie di proprietà riconducibili al comportamento della distribuzione campionaria di T_n (ovvero del modello di probabilità che può essere usato per T_n)

PROBLEMI DI STIMA

proprietà degli stimatori ⇝ riconducibili a proprietà della distribuzione campionaria di T_n

POSIZIONE

non distorsione

riassume il comportamento medio della distribuzione campionaria di T_n

VARIABILITA'

efficienza

riassume la distribuzione campionaria di T_n in termini di dispersione rispetto alla media

FORMA

"NORMALITA"

molti stimatori possono essere descritti o approssimati da una legge normale

PROBLEMI DI STIMA

POPOLAZIONE

Stima di μ — $X \sim N(\mu, \sigma^2)$ $\Rightarrow$ $\bar{X}_n \sim N\left(\mu, \frac{\sigma^2}{n}\right)$

Stima di μ — $X \sim f$ $\Rightarrow$ $\bar{X}_n \xrightarrow[n \geq 30]{} N\left(\mu, \frac{\sigma^2}{n}\right)$

stimatore $\bar{X}_n = \frac{\sum X_i}{n}$

Stima di π — $X \sim Ber(\pi)$ — la distribuzione esatta di $\hat{\pi}$ può essere ottenuta facendo riferimento alla binomiale

stimatore $T_n = \hat{\pi} = \bar{X}_n = \frac{\sum X_i}{n} \sim bin(n, \pi)$

sfruttando il TLC, se $n \geq 30$ si ha:

$$\hat{\pi} \longrightarrow N\left(\pi, \frac{\pi(1-\pi)}{n}\right)$$

Stima di σ^2 → lo stimatore "naturale" sarebbe:

$$\tilde{\sigma}^2 = \frac{\sum_{i=1}^{n} (x_i - \mu)^2}{n}$$

→ questo richiederebbe la conoscenza di μ (poco plausibile)

→ si può stimare μ usando $\bar{x}_n$

STIMATORE VARIANZA CAMPIONARIA

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

per correggere un (piccolo) problema di distorsione di $\hat{\sigma}^2$ si divide per $n-1$ (vedi pagine più avanti per i dettagli tecnici)

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

STIMATORE VARIANZA CAMPIONARIA CORRETTA

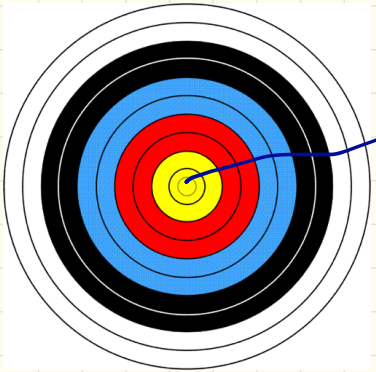
La distribuzione campionaria di s^2 richiede l'introduzione di un nuovo modello di probabilità: la v.c. X^2 → caratterizzata da un unico parametro, detto gradi di libertà (g.d.l.)

↓

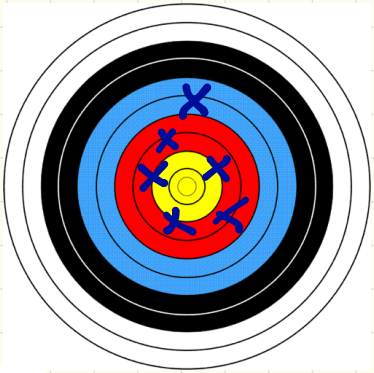
è definita dalla somma di n v.c. normali standardizzate al quadrato

↓

nel caso dell'inferenza su varianza è necessaria l'ipotesi di normalità (rispetto ai problemi relativi alla media o alla proporzione, l'inferenza sulla varianza è più sensibile a violazioni dell'ipotesi di normalità)



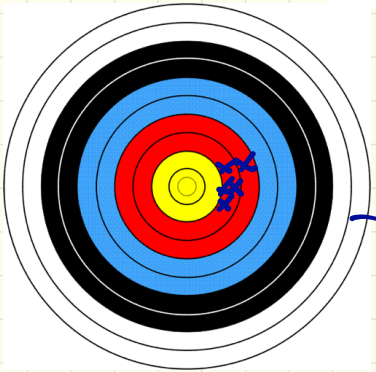
il centro del bersaglio è θ
(l'obiettivo finale)



comportamento medio

$$E(T_m) = \theta$$

T_m è non distorto per θ



$E(T_m) \neq \theta$

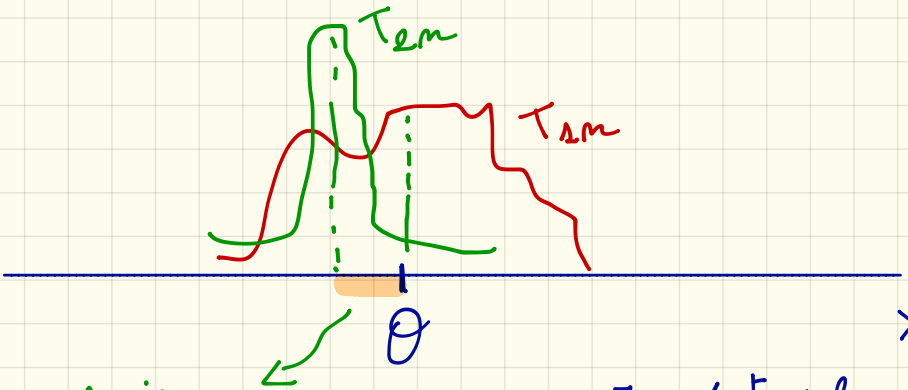
qui T_m è distorto: $E(T_m) - \theta =$

$= d(T_m) = b(T_m)$

distorsione (bias)

Se $b(T_m) < 0 \rightsquigarrow$ sottostima di θ

$b(T_m) > 0 \rightsquigarrow$ sovrastima di θ



misura della distorsione di T_{em}

$\times$

In sintesi la distorsione sintetizza $E(T_m)$

$\downarrow$

posizione della v.c. stimatore

$$\bar{X}_n = \frac{\sum X_i}{n}$$

↓

MEMO: vedi lezione sulle combinazioni lineari di v.c.

$$E(\bar{X}_n) = \text{combinaz. lineare di } n \text{ v.c. i.i.d.} = \mu$$

↓

la media campionaria è non distorta per $\theta = \mu$

$F_n = \bar{X}_n$ particolare media di osservazioni 0, 1

↓

estratte da una popolazione $X \sim Ber(\pi)$

↓

$E(F_n) = \mu = \pi$

→ F_n è uno stimatore non distorto per π } $F_n = \hat{\pi}$ (altro possibile nome)

Immaginiamo di essere interessati a stimare σ^2

↓

il nostro θ

Sulla popolazione

$$\sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}$$

utilizzando il principio di naturalità

$$\frac{\sum_{i=1}^{n} (x_i - \mu)^2}{n}$$

μ è un parametro di disturbo (non è il vero obiettivo dell'inferenza)

è poco plausibile che io abbia informazioni su μ

↓

varianza campionaria

$$\hat{\sigma}^2 = \frac{\sum (x_i - \bar{x}_n)^2}{n}$$

utilizzo la stima di μ

$E(\hat{\sigma}^2) \neq \sigma^2$ esempio di stimatore distorto

$$E[\hat{\sigma}^2] = E\left[\frac{\sum_i (X_i - \bar{X}_m)^2}{m}\right] = E\left[\frac{\sum X_i^2}{m} - \bar{X}_m^2\right] =$$

(proprietà varianza)

$$= E\left[\frac{\sum X_i^2}{m}\right] - E(\bar{X}_m^2) =$$

$$\frac{\sum E(X_i^2)}{m}$$

NOTA: $Var(\bar{X}_m) = E[\bar{X}_m - E(\bar{X}_m)]^2$

$$E(\bar{X}_m^2) - [E(\bar{X}_m)]^2$$

$$E(\bar{X}_m^2) = Var(\bar{X}_m) + [E(\bar{X}_m)]^2 =$$

$$= \frac{\sigma^2}{m} + \mu^2$$

$$E(\hat{\sigma}^2) = \frac{\sum_i E(X_i^2)}{n} - \left(\frac{\sigma^2}{n} + \mu^2\right) =$$

$$= \frac{\sum_i E(X_i^2)}{n} - \mu^2 - \frac{\sigma^2}{n} = \sigma^2 - \frac{\sigma^2}{n} =$$

$$= \frac{n\sigma^2 - \sigma^2}{n} =$$

$$= \frac{n-1}{n}\sigma^2$$

$$Var(X_i) = Var(X)$$

$\hat{\sigma}^2$ è distorto per campioni finiti ma asintoticamente non distorto

$n \to \infty$ allora $\frac{n-1}{n} \to 0$

← [la distorsione diminuisce al crescere di n

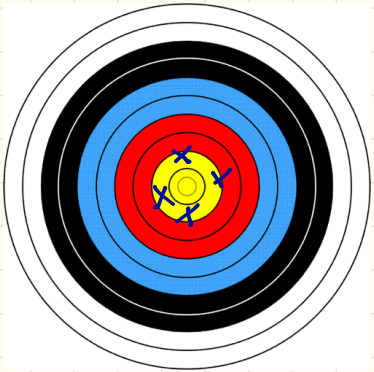
In questo caso è possibile correggere la distorsione:

$$T_n = \frac{n}{n-1}\,\hat{\sigma}^2 = \frac{\cancel{n}}{n-1}\,\frac{\sum (x_i - \bar{x}_n)^2}{\cancel{n}} = S^2$$

varianza campionaria corretta

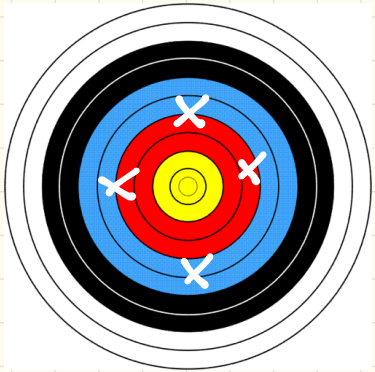
$$E(S^2) = E\left(\frac{n}{n-1}\,\hat{\sigma}^2\right) = \frac{n}{n-1}\,E(\hat{\sigma}^2) =$$

$$= \frac{n}{n-1}\,\frac{n-1}{n}\,\sigma^2 = \sigma^2$$



T_{1m} e T_{2m} siano due stimatori non distorti per θ

$$E(T_{1m}) = E(T_{2m}) = \theta$$



$$Var(T_{1m}) = E[T_{1m} - E(T_{1m})]^2 = E[T_{1m} - \theta]^2$$

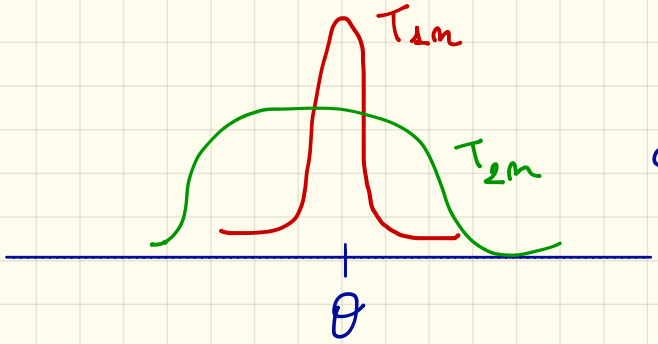
$$Var(T_{2m}) = E[T_{2m} - \theta]^2$$

$Var(T_{1m}) \leq Var(T_{2m})$

T_{1m} è preferibile

sono equivalenti

T_{1m} è più efficiente di T_{2m} (effic. in senso relativo rispetto a T_{2m})

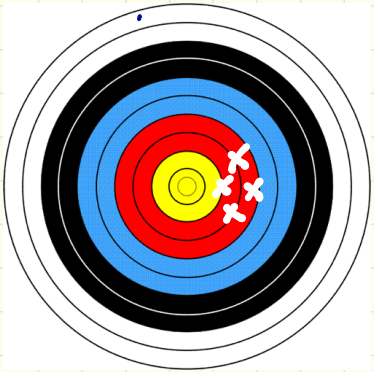


Efficienza in senso relativo consiste in un confronto tra le misure di variabilità delle distribuzioni campionarie dei due stimatori

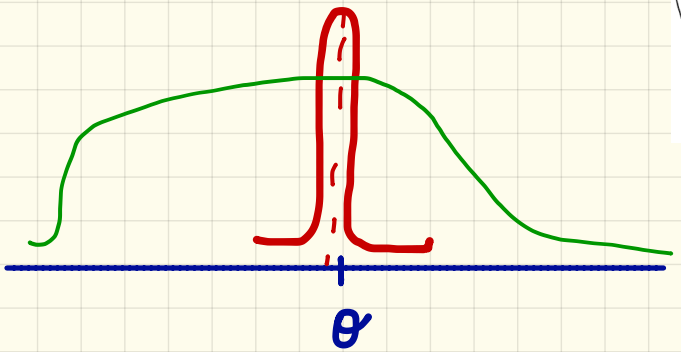
Efficienza in senso assoluto → permette di individuare un valore limite per la varianza di uno stimatore per un dato parametro al disotto del quale non si può scendere

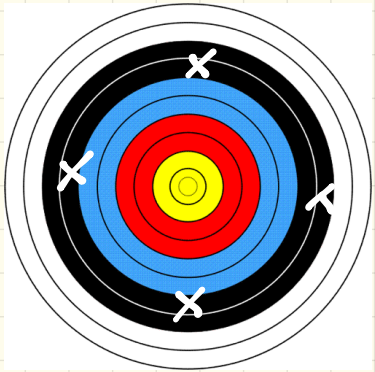
limite di Cramér - Rao

Se $Var(T_n)$ = valore limite] T_n è efficiente in senso assoluto



T_{1m}





T_{2m}

Si introduce un indice: errore quadratico medio EQM
(mean squared error MSE)

che tiene conto contemporaneamente delle due proprietà

$$MSE(T_m) = E\left[(T_m - \theta)^2\right] = Var(T_m) + b^2(T_m)$$

$$MSE(T_m) = E[(T_m - \theta)^2] =$$

$$= E[(T_m - E(T_m) + E(T_m) - \theta)^2] =$$

$$= E\{([T_m - E(T_m)] + [E(T_m) - \theta])^2\} =$$

$$= \underbrace{E[T_m - E(T_m)]^2}_{Var(T_m)} + E[\underbrace{E(T_m) - \theta}_{\text{costante}}]^2 + 2E([T_m - E(T_m)][\underbrace{E(T_m) - \theta}_{\text{costante}}])$$

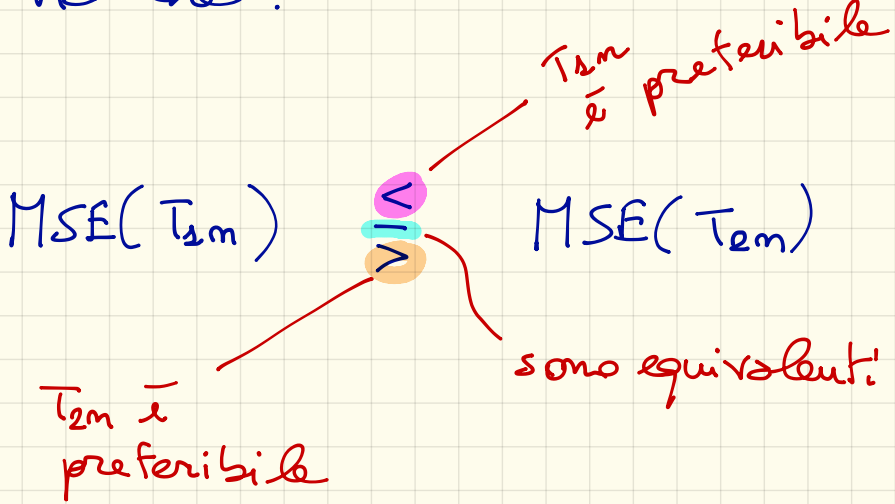
$= 0$ (variabile scarto)

$E(\text{costante}) = \text{costante}$

$$[\underbrace{E(T_m) - \theta}]^2$$

$b^2(T_m)$

Dati due stimatori T_{1m} e T_{2m} (non necessariamente non distorti) si ha che:



Esempio di scelta tra due stimatori alternativi

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

- N → popolazione
- μ → obiettivo stima
- popolazione simmetrica → $\mu = Me$

$\mu = Me$:

- $\bar{X}_n$ media campionaria
- $\hat{Me}$ mediana campionaria

non distorti $\rightsquigarrow$ $E(\bar{X}_n) = E(\hat{Me}) = \mu$

$$Var(\bar{X}_n) = \frac{\sigma^2}{n} < \frac{\pi \sigma^2}{2n} = Var(\hat{Me})$$

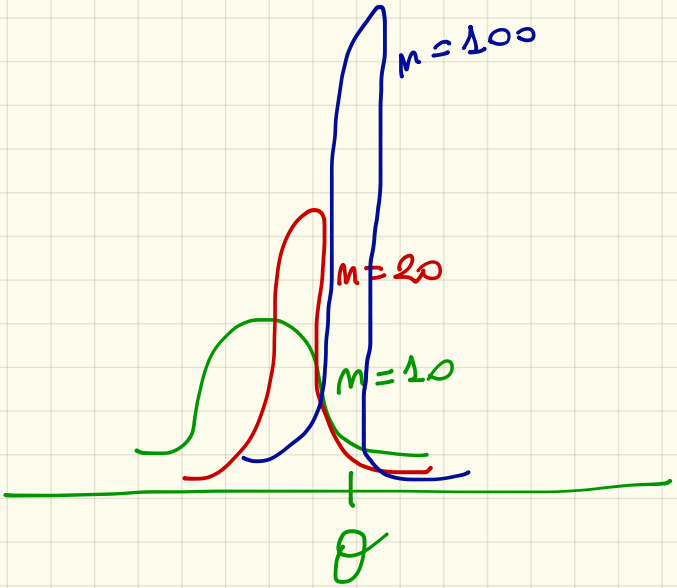
$$\frac{Var(\hat{Me})}{Var(\bar{X}_n)} = \frac{\pi \sigma^2 / 2n}{\sigma^2 / n} = \frac{\pi}{2} \approx 1.57$$

Le proprietà degli stimatori possono essere studiate sia su campioni finiti (PROPRIETA' FINITE) che al crescere della numerosità campionaria (PROPRIETA' ASINTOTICHE)

$E(\hat{\sigma}^2) \neq \sigma^2$ distorto ⟶ diventa non distorto asintoticamente

Tra le differenti proprietà asintotiche (che studiano il comportamento della distribuzione campionaria quando $n \to +\infty$) vale la pena richiamare la proprietà della consistenza in media quadratica

↳ $\lim [E(T_n - \theta)^2] = 0$



La consistenza in media quadratica vale se e solo se:

① $\lim_{n\to\infty} E(T_n) = \theta$

T_n è asintoticamente non distorto

② $\lim_{n\to\infty} Var(T_n) = 0$

Una proprietà utile relativa alla forma della distribuzione campionaria di uno stimatore T_n è la normalità asintotica:

$$\lim_{n\to\infty} \frac{T_n - E(T_n)}{\sqrt{Var(T_n)}} \longrightarrow N(0,1)$$

RIEPILOGO DISTRIBUZIONI CAMPIONARIE
CASO DI INFERENZA SU UNA SOLA POPOLAZIONE

Schema sintesi (lessico di base)

$X \sim f(x;\theta)$ → parametro o vettore di parametri

modello popolazione

ci limitiamo allo studio di un parametro alla volta

$X_1, X_2, \dots, X_n$

campione casuale n-pla di v.c.

le X_i sono i.i.d. $\sim X$

ipotesi di campionamento con ripetizione

$x_1, x_2, \dots, x_n$

campione osservato

(unica) realizzazione disponibile del campione casuale

l'inferenza lavora su questo unico campione

La scelta dei metodi da usare si basa sul

PRINCIPIO DEL CAMPIONAMENTO RIPETUTO

ovvero lo studio delle statistiche di interesi campionarie su tutti i "possibili" campioni ↝ la scelta della statistica (stimatore) da usare si basa sulle caratteristiche della corrispondente distribuzione campionaria

$$T(X_1, X_2, \dots, X_n) = T_n \qquad \text{stimatore (v.c.)}$$

$$T(x_1, x_2, \dots, x_n) = t_n \qquad \text{stima}$$

NOTA: $\sqrt{\text{Var}(T_n)}$ (lo scarto quadratico medio) viene talora chiamato STANDARD ERROR, ERRORE STANDARD o DEVIAZIONE STANDARD

INFERENZA SULLA MEDIA μ DI UNA POPOLAZIONE

① $X \sim N(\mu, \sigma^2)$

θ

σ^2 → nota

$\bar{X}_n$

$E(\bar{X}_n) = \mu$

$Var(\bar{X}_n) = \frac{\sigma^2}{n}$

consistente → raggiunge il limite di Cramer-Rao (efficiente in senso assoluto)

(?) come si distribuisce (?)

proprietà riproduttiva → se $X \sim N$ allora $\bar{X}_n \sim N$

distribuzione esatta

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

(2)

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

θ

σ^2 → incognito [PARAMETRO DI DISTURBO]

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

σ → S

uso il campione per stimare anche il parametro di disturbo

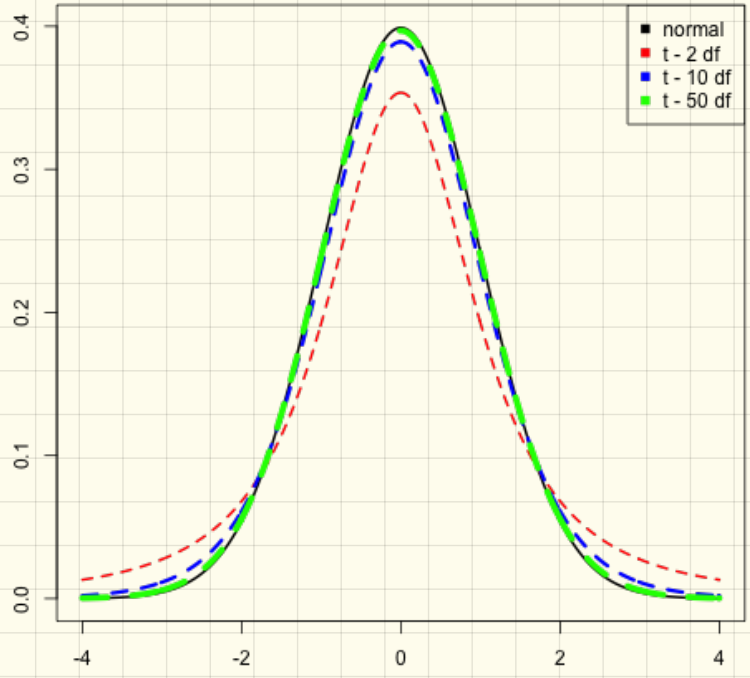
t_{n-1}

gradi di libertà

unico parametro t di Student

al crescere dei g. d. l., la v. c. t converge ad una normale standardizzata

esistono delle tavole che contengono alcuni percentili di più comune utilizzo di questa distribuzione



(3) $X \sim f$ — posso conoscerla e so che non è normale

non la conosco

sono comunque interessato a stimarne la μ sulla base delle osservazioni campionarie

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{\text{converge}} N(0,1)$$ di distribuz. asintotica

$$\frac{\bar{X}_n - \mu}{S/\sqrt{n}} \to t_{n-1} \to N(0,1)$$

se n cresce

INFERENZA SULLA VARIANZA σ^2

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

θ

di solito non ho informazioni su μ: parametro di disturbo

Sotto questa ipotesi si può mostrare

$$\frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{n-1}$$

unico parametro (gradi di libertà)

l'inferenza su σ^2 è meno robusta rispetto alla validazione dell'ipotesi di normalità

INFERENZA SULLA PROPORZIONE π

$X \sim Ber(\pi)$

↳ in questo caso la distribuzione esatta di $T_n = \bar{X}_n = \hat{\pi}$ può essere ricavata sfruttando la distribuzione binomiale

↓

se n è "abbastanza" grande si può anche in questo caso sfruttare il teorema del limite centrale per ricavare la distribuzione asintotica

$$\frac{\hat{\pi} - E(\hat{\pi})}{\sqrt{Var(\hat{\pi})}} = \frac{\hat{\pi} - \pi}{\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}} \xrightarrow{TLC} N(0,1)$$

	Task Description	Due Date	Status
☐	RIEPILOGO DISTRIBUZIONI CAMPIONARIE CASO DI INFERENZA SU DUE POPOLAZIONI		
☐	Inferenza sul confronto tra le medie di due popolazioni: caso di varianze note e caso di varianze incognite ma supposte uguali, rimozione dell'ipotesi di normalità		
☐	Inferenza sul confronto tra le varianze di due popolazioni normali		
☐	Inferenza sul confronto tra le proporzioni di due popolazioni bernoulliane		
☐			
☐			
☐			
☐			
☐			
☐			

In molti casi si può essere interessati a confrontare i parametri di due popolazioni. In particolare:

- confronto tra le due medie (caso di varianze note e caso di varianze incognite) di due popolazioni normali

  ↓

  questo schema, sfruttando il TLC, può essere esteso anche al caso di due generiche popolazioni se le numerosità campionarie sono adeguate

- confronto tra le varianze di due popolazioni normali
- confronto tra le proporzioni di due popolazioni di Bernoulli

CONFRONTO TRA LE MEDIE DI DUE POPOLAZIONI NORMALI (caso di varianze note)

Siano X e Y due v.c. normali ed indipendenti

$X \sim N(\mu_X, \sigma_X^2)$ — σ_X^2: nota a priori

$Y \sim N(\mu_Y, \sigma_Y^2)$ — σ_Y^2: nota a priori

$\Downarrow$

$\bar{X}_n \sim N\left(\mu_X, \frac{\sigma_X^2}{n_X}\right)$ $\rightsquigarrow$ per la proprietà riproduttiva della normale $\leftsquigarrow$ $\bar{Y}_n \sim N\left(\mu_Y, \frac{\sigma_Y^2}{n_Y}\right)$

$\Downarrow$ se si considera la v.c. $\bar{X}_n - \bar{Y}_n$ (teorema sulle v.c. normali)

$\Downarrow$ qualsiasi combinazione lineare di v.c. normali è ancora normale

$$\bar{X}_n - \bar{Y}_n \sim N\left(\mu_X - \mu_Y, \frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{n}\right)$$

$$\frac{(\bar{X}_n - \bar{Y}_n) - (\mu_X - \mu_Y)}{\sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}}} \sim N(0,1)$$

CONFRONTO TRA LE MEDIE DI DUE POPOLAZIONI NORMALI (caso di varianze incognite)

Se nel precedente schema teorico le due varianze σ_X^2 e σ_Y^2 sono incognite si parla di problema di Behrens-Fisher

↓

come nel caso dell'inferenza su una sola popolazione, quando si sostituisce alla varianza la sua stima corretta, la distribuzione campionaria dello stimatore passa da una distribuzione normale ad una t di Student

↓

le formule per il calcolo dei g.d.l. della t di Student risultante sono molto complesse

↓

il caso più semplice da gestire si basa sull'ipotesi di OMOSCHEDASTICITÀ: le due popolazioni hanno varianza incognita che però si può assumere uguale → $\sigma_X^2 = \sigma_Y^2 = \sigma^2$

Si ha allora:

$$X \sim N(\mu_X, \sigma^2) \quad \Rightarrow \quad \bar{X}_n \sim N\left(\mu_X, \frac{\sigma^2}{n_X}\right)$$

$$Y \sim N(\mu_Y, \sigma^2) \quad \Rightarrow \quad \bar{Y}_n \sim N\left(\mu_Y, \frac{\sigma^2}{n_Y}\right)$$

$$\frac{(\bar{X}_n - \bar{Y}_n) - (\mu_X - \mu_Y)}{\sqrt{\frac{\sigma^2}{n_X} + \frac{\sigma^2}{n_Y}}} = \frac{(\bar{X}_n - \bar{Y}_n) - (\mu_X - \mu_Y)}{\sqrt{\sigma^2\left(\frac{1}{n_X} + \frac{1}{n_Y}\right)}}$$

sotto l'ipotesi di omoschedasticità ho un solo parametro di disturbo

posso stimarlo usando lo stimatore congiunto s_p^2

$$\rightarrow \frac{(\bar{X}_n - \bar{Y}_n) - (\mu_X - \mu_Y)}{\sqrt{s_p^2\left(\frac{1}{n_X} + \frac{1}{n_Y}\right)}} \sim t_{n_X + n_Y - 2}$$

p sta per pooled (congiunto)

$$s_p^2 = \frac{(n_x - 1)\, s_x^2 + (n_y - 1)\, s_y^2}{n_x + n_y - 2}$$

lo stimatore congiunto della varianza comune è una media ponderata delle due varianze campionarie con pesi date dalle rispettive numerosità

esplicitando le formule per s_x^2 e s_y^2, si ottiene:

$$s_p^2 = \frac{(n_x - 1)\dfrac{\sum_i (x_i - \bar{x}_{n_x})^2}{(n_x - 1)} + (n_y - 1)\dfrac{\sum_i (y_i - \bar{y}_{n_y})^2}{(n_y - 1)}}{(n_x + n_y - 2)} =$$

$$= \frac{\sum_i (x_i - \bar{x}_{n_x})^2 + \sum_i (y_i - \bar{y}_{n_y})^2}{n_x + n_y - 2} = \frac{Dev(X) + Dev(Y)}{n_x + n_y - 2}$$

lo stimatore congiunto di σ^2 è la somma delle due devianze campionarie

$X \sim f$ $\quad$ rimuoviamo l'ipotesi di normalità delle due popolazioni $\quad$ $Y \sim f$

$\downarrow$ $\quad$ $\downarrow$

$\bar{X}_m \not\sim N\left(\mu_x, \frac{\sigma_x^2}{m_x}\right)$ $\quad$ $\bar{Y}_m \not\sim N\left(\mu_y, \frac{\sigma_y^2}{m_y}\right)$

$\rightarrow$ $(m_x \geq 30)$ $\quad$ $\rightarrow$ $(m_y \geq 30)$

$$\bar{X}_m - \bar{Y}_m \not\sim N\left(\mu_x - \mu_y, \frac{\sigma_x^2}{m_x} + \frac{\sigma_y^2}{m_y}\right) \rightarrow N$$

posso usare ancora le stesse statistiche

robuste rispetto all'ipotesi di violazione di normalità

anche in questo caso, nel caso di varianze incognite, per semplicità lavoriamo sotto l'ipotesi di omoschedasticità

CONFRONTO TRA LE VARIANZE DI DUE POPOLAZIONI NORMALI

$X \sim N(\mu_X, \sigma_X^2)$

$Y \sim N(\mu_Y, \sigma_Y^2)$

$$\frac{(n_X - 1)\, S_X^2}{\sigma_X^2} \sim \chi^2_{n_X - 1}$$

$$\frac{(n_Y - 1)\, S_Y^2}{\sigma_Y^2} \sim \chi^2_{n_Y - 1}$$

$$\frac{\dfrac{(n_X - 1)\, S_X^2}{\sigma_X^2} \Big/ (n_X - 1)}{\dfrac{(n_Y - 1)\, S_Y^2}{\sigma_Y^2} \Big/ (n_Y - 1)} = \frac{S_X^2 / \sigma_X^2}{S_Y^2 / \sigma_Y^2} \sim F_{n_X - 1,\, n_Y - 1}$$

NOTA: il rapporto tra due v.c. χ^2, ciascuna divisa per i corrispondenti gdl è una v.c. F di Fisher

due parametri

gdl del numeratore

gdl del denominatore

anche in questo caso è disponibile una tavola con i percentili di più comune utilizzo

NOTA TECNICA: mentre nel caso dell'inferenza sul confronto tra le due medie, nel caso di numerosità campionarie adeguate si può utilizzare comunque lo stesso schema anche se le due popolazioni non sono normali, nel caso di inferenza sulle due varianze la procedura risulta abbastanza sensibile a violazioni dell'ipotesi di normalità

CONFRONTO TRA LE PROPORZIONI DI DUE POPOLAZIONI BERNOULLIANE

$X \sim Ber(\pi_X)$ $\quad$ $Y \sim Ber(\pi_Y)$

la distribuzione esatta di $\mathcal{F}_X = \bar{X}_m = \hat{\pi}_X$ è ricavabile a partire dalla v.c. binomiale

lo stesso per $\mathcal{F}_Y = \bar{Y}_m = \hat{\pi}_Y$

se i due campioni sono sufficientemente grandi si può sfruttare il TLC per ottenere le due distribuzioni campionarie

$$(\hat{\pi}_X - \hat{\pi}_Y) \xrightarrow{TLC} N\left(\pi_X - \pi_Y, \frac{\pi_X(1-\pi_X)}{m_X} + \frac{\pi_Y(1-\pi_Y)}{m_Y}\right)$$

la varianza congiunta può essere stimata usando lo stimatore $\hat{\pi}_P$

$$\hat{\pi} = \frac{\sum X_i + \sum Y_i}{m_X + m_Y}$$ proporzione di successi nei due campioni

$$\frac{(\hat{\pi}_X - \hat{\pi}_Y) - (\pi_X - \pi_Y)}{\sqrt{\hat{\pi}_p(1-\hat{\pi}_p)\left(\frac{1}{n_X}+\frac{1}{n_Y}\right)}} \xrightarrow{TLC} N(0,1)$$

dove: $\hat{\pi}_p = \dfrac{\sum_{i=1}^{M_X} X_i + \sum_{i=1}^{M_Y} Y_i}{M_X + M_Y}$