

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 r.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 $\leadsto$ insieme dati POPOLAZIONE

↓
obiettivo sintesi

Parte inferenziale $\leadsto$ sottoinsieme della POPOLAZIONE

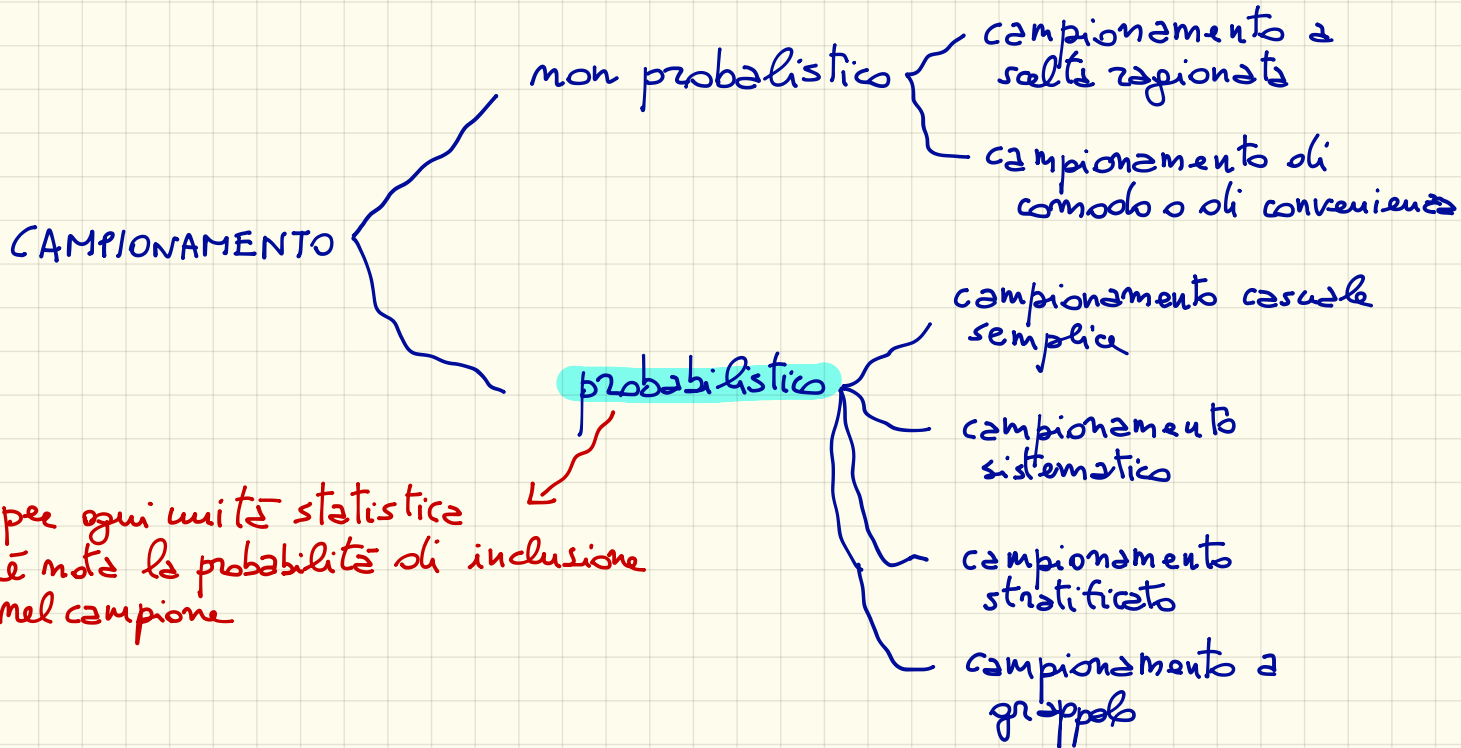
se il campione è estratto casualmente gli strumenti della probabilità ci aiutano in questo passaggio inverso

↓
INFERENZA

campione

↓
sintesi del campione

↑
obiettivo rivolto ad ottenere informazioni sulla popolazione



campionamento casuale
semplice

con ripetizione (più "naturale")

tutte le unità
statistiche della
popolazione hanno la
stessa probabilità di
inclusione nel campione

senza ripetizione (vantaggio:
semplicità)

↓
se $N \rightarrow \infty$ 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 $\rightarrow$ 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
- Acto

3 livelli o strati corrispondono
a urne 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 $\sigma^2 = \text{Var}(X) = \text{Var}(X_i)$.

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

$$W = \sum_i a_i X_i$$

Si ha:

$$\mu_W = E(W) = \mu \sum_i a_i$$

$$\sigma_W^2 = \text{Var}(W) = \sigma^2 \sum_i a_i^2$$

Due particolari combinazioni lineari:

• somma ($a_i = 1 \ \forall i$) $\rightarrow W = \sum_i X_i \left\{ \begin{array}{l} \mu_W = n\mu \\ \sigma_W^2 = n\sigma^2 \end{array} \right.$

• media ($a_i = \frac{1}{n} \ \forall i$) $\rightarrow W = \frac{\sum_i X_i}{n} = \bar{X}_n \left\{ \begin{array}{l} \mu_W = \mu \\ \sigma_W^2 = \frac{\sigma^2}{n} \end{array} \right.$

Cosa è possibile dire circa la distribuzione di W ?

Se $X \sim \text{Bec}(\pi)$ $\rightarrow W = \sum X_i \sim \text{bin}(n, \pi)$ $\left\{ \begin{array}{l} E(W) = n\pi \\ \text{Var}(W) = n\pi(1-\pi) \end{array} \right.$

$\rightarrow W = \frac{\sum X_i}{n}$ trasformazione di una
v.c. binomiale

$$\left\{ \begin{array}{l} E(W) = \pi \\ \text{Var}(W) = \frac{\pi(1-\pi)}{n} \end{array} \right.$$

Se $X \sim N(\mu, \sigma^2) \rightarrow W = \sum a_i X_i \sim N\left(\mu \sum a_i, \sigma^2 \sum a_i^2\right)$

proprietà riproduttiva v.c. normale
(distribuzione esatta)

Se $X \sim f(x; \theta) \rightarrow W = \sum a_i X_i \xrightarrow{n \rightarrow \infty} N\left(\mu \sum a_i, \sigma^2 \sum a_i^2\right)$

teorema del limite centrale
(distribuzione approssimata)

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

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

ATTENZIONE: $\rightarrow$
tutta l'inferenza si
basa su questa ipotesi

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

$\rightarrow$ parametro di interesse

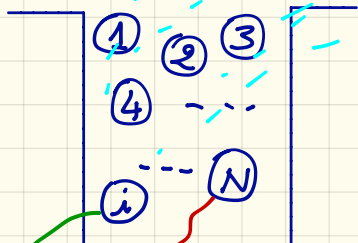
$\downarrow$
l'inferenza parametrica mira ad
ottenere informazioni su θ a partire
dei dati campionari

$\nwarrow$
(H)

Spazio parametrico



insieme dei possibili valori che può assumere
il parametro θ



N : numerosità della popolazione
 ($N \rightarrow \infty$: popolazioni infinite)

[l' i -esima unità della popolazione]

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

CAMPIONE OSSERVATO

$x_1, x_2, \dots, x_i, \dots, x_n$

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

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

campione osservato (m valori)

$(X_1, X_2, \dots, X_i, \dots, X_m)$

campione casuale (m r.c.)

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. $\sim X$

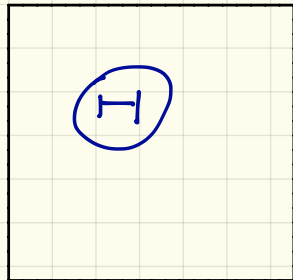
parallelo con le combinazioni lineari
di r.c. richiamato ad inizio lezione

l'insieme di tutti i possibili
campioni di ampiezza m
(ovvero l'insieme di tutti i possibili
valori che possono comporre il campione
osservato) si chiama spazio campionario
e si indica con $\mathcal{X}_m$

NOTA: da non confondere
con Ω (spazio campione)

GLI "SPAZI" DELL'INFERENZA

SPAZIO
PARAMETRICO



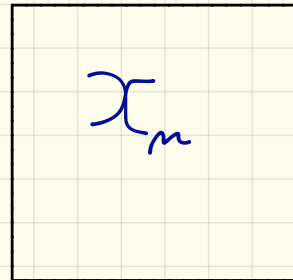
↓
lo spazio di interesse
↓
rispetto cui devo "decidere"

SPAZIO DELLA
STATISTICA



↑
in molti
problemi è
di dimensione
univariata

SPAZIO
CAMPIONARIO



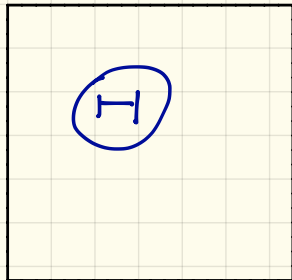
↓
lo spazio di
lavoro
↓
sul quale effettivamente
"decisione"

di solito non si
lavora direttamente su
 x_m ma su una sua opportuna
sintesi

→ STATISTICA

GLI "SPAZI" DELL'INFERENZA

SPAZIO
PARAMETRICO

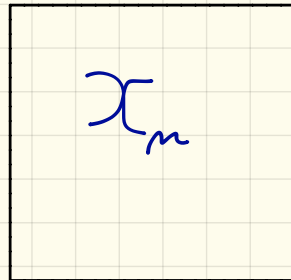


SPAZIO DELLA
STATISTICA



la distribuzione di
probabilità di T_m si chiama
distribuzione campionaria

SPAZIO
CAMPIONARIO



e che quindi
può essere
calcolata

qualsunque funzione
del campione casuale che
non dipende da parametri incogniti

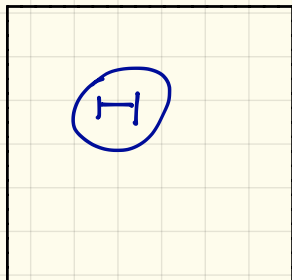
$$T_m = T(X_1, X_2, \dots, X_i, \dots, X_n)$$

è una v.c.
(funzione di)
 n v.c.

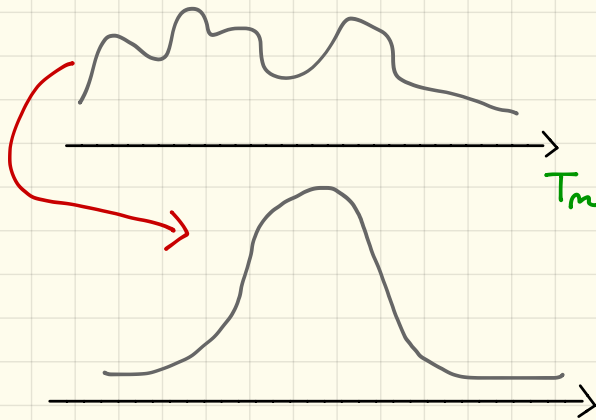
$$t_m = T(x_1, x_2, \dots, x_i, \dots, x_n)$$

valore assunto
dalla statistica in
corrispondenza del
campione osservato

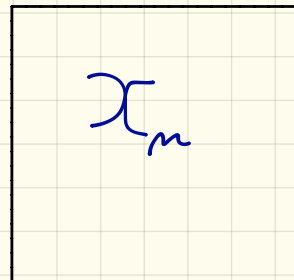
SPAZIO
PARAMETRICO



SPAZIO DELLA
STATISTICA



SPAZIO
CAMPIONARIO



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

$$X \sim \text{Bee}(\pi)$$

$$\bar{x}_m = \hat{\pi} = \frac{\sum x_i}{n} \rightarrow N\left(\pi, \frac{\pi(1-\pi)}{n}\right)$$

distribuzione approssimata
(TLC)

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

$$\bar{x}_m = \frac{\sum x_i}{n} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

distrib. esatta (proprietà riproduttiva normale)

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

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

distribuzione
approssimata (TLC)

SULLA SCELTA DELLA "MIGLIORE" STATISTICA T_m



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_m)

→ problemi di stima → T_m si chiama stimatore
→ verifiche di ipotesi → T_m si chiama statistica test



due requisiti essenziali:

- **maternalità** applico la stessa statistica che applicerei sulle popolazioni all'unico campione disponibile
- **SUFFICIENZA** → il requisito essenziale è che T_m sia in grado di catturare tutte le informazioni utili su θ contenute nel campione

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

Inferenza su μ [DISTRIBUZIONE MEDIA CAMPIONARIA]

↓

$$T_m = \bar{X}_m = \frac{\sum X_i}{m} \quad \text{si usa la statistica media campionaria}$$

• se $X \sim N(\mu, \sigma^2) \Rightarrow \bar{X}_m \sim N\left(\mu, \frac{\sigma^2}{m}\right)$ proprietà riproduttiva r.v. normale

• se $X \sim f(x; \theta) \Rightarrow \bar{X}_m \xrightarrow{m \rightarrow \infty} N\left(\mu, \frac{\sigma^2}{m}\right)$ teorema del limite centrale

↓
X non è normale
oppure non è possibile
assumere la normalità

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

Inferenza su π [DISTRIBUZIONE PROPORZIONE CAMPIONARIA]



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

$\left(\begin{array}{l} \frac{\sum X_i}{n} \text{ si tratta di una media di osservazioni} \\ \{0, 1\} \end{array} \right)$
↳ possesso della caratteristica di interesse (successo)

In questo caso: $X \sim \text{Ber}(\pi)$

↳ proporzione nella popolazione

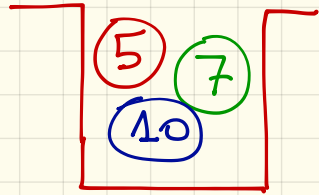
$\hat{J}_n = \hat{\pi}$ è una trasformazione di una v.c. binomiale

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

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

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

CASO STUDIO (ci permette di introdurre i concetti basilari)



popolazione



X n. errori per
pagine di tre
dattilografi

$N = 3$

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
 $n = 2$

① quanti sono i possibili campioni che ②
potrei osservare

$$3 \times 2 = 6$$

estrazione senza
rimessa

$$3 \times 3 = 9$$

estraz. con
rimessa

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

$X_2 \leadsto$ 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 rimesse

$X_1 (= X_2)$	senza rimesse	con rimesse
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 rimesse 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

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

fenomeno di interesse
popolazione

nell'es. $\bar{x}$ il m. di errori
commesso dalle dattilografe

modello di probabilità che si
pensa possa descrivere adeguatamente
il fenomeno

tutto quello che vorremo è seguire
si basa sull'ipotesi che $h \approx f$ sia
adeguata

parametro o vettore
di parametri che
caratterizzano f

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
maturità



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

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

Immaginiamo di essere interessati a studiare la
media della popolazione

Task Description

Due Date

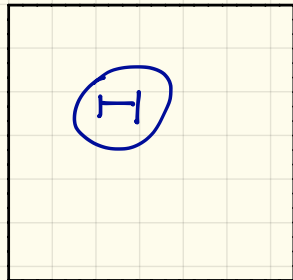
Status

☐	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 <u>UNA SOLA</u> POPOLAZIONE		
☐	Inferenza sulla <u>media</u> : caso di varianza nota, di varianza incognita e rimozione dell'ipotesi di normalità		
☐	Inferenza sulla <u>varianza</u> di una popolazione normale		
☐	Inferenza sulla <u>proporzione</u> di una popolazione bernoulliana		

GLI "SPAZI" DELL'INFERENZA

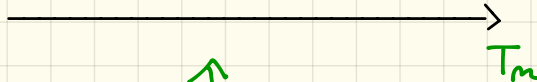
MEMO

SPAZIO
PARAMETRICO



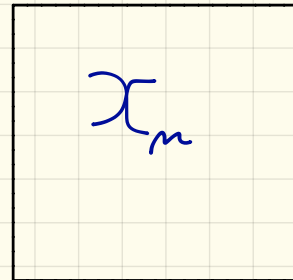
↓
lo spazio di interesse
↓
rispetto cui devo "decidere"

SPAZIO DELLA
STATISTICA



↑
in molti
problemi è
di dimensione
univariata

SPAZIO
CAMPIONARIO



↓
lo spazio di
lavoro
↓
sul quale effettivamente
"decisione"

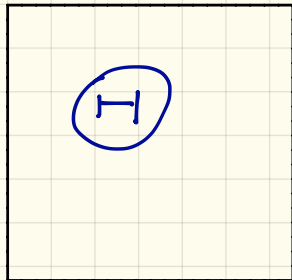
di solito non si
lavora direttamente su
 x_m ma su una sua opportuna
sintesi

→ STATISTICA

GLI "SPAZI" DELL'INFERENZA

MEMO

SPAZIO
PARAMETRICO

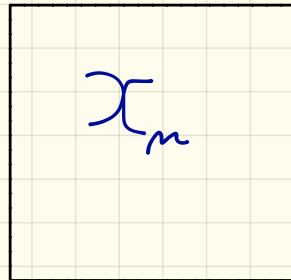


SPAZIO DELLA
STATISTICA



la distribuzione di
probabilità di T_m si chiama
distribuzione campionaria

SPAZIO
CAMPIONARIO



e che quindi
può essere
calcolata

qualsunque funzione
del campione casuale che
non dipende da parametri incogniti

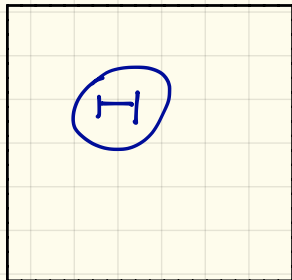
$$T_m = T(X_1, X_2, \dots, X_i, \dots, X_n)$$

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

è una v.c.
(funzione di)
n v.c.

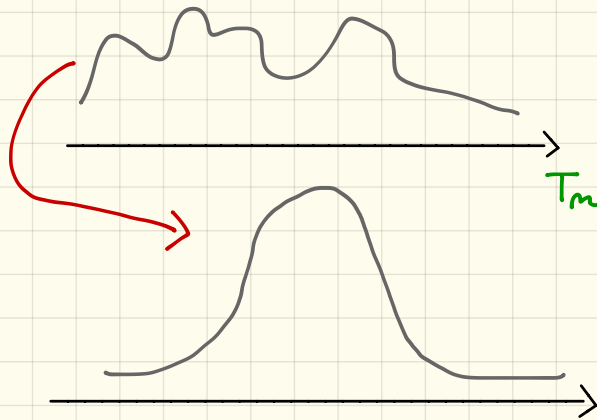
valore assunto
dalla statistica in
corrispondenza del
campione osservato

SPAZIO
PARAMETRICO

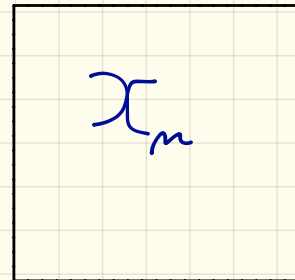


MEMO

SPAZIO DELLA
STATISTICA



SPAZIO
CAMPIONARIO



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

$$X \sim \text{Bee}(\pi)$$

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

distribuzione approssimata
(TLC)

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

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

distrib. esatte (proprietà riproduttiva normale)

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

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

distribuzione
approssimata (TLC)

SULLA SCELTA DELLA "MIGLIORE" STATISTICA T_m

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_m)

→ problemi di stima → T_m si chiama stimatore

→ verifiche di ipotesi → T_m si chiama statistica test

due requisiti essenziali:

- **maternalità** applico la stessa statistica che applicerei sulle popolazioni all'unico campione disponibile
- **SUFFICIENZA** → il requisito essenziale è che T_m sia in grado di catturare tutte le informazioni utili su θ contenute nel campione

una volta calcolata la statistica T_m 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

verifiche di ipotesi



si vuole verificare la
validità di un'ipotesi
su θ a partire da 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 proba-
bilità che può essere usato per T_n)

PROBLEMI DI STIMA



proprietà degli stimatori $\leadsto$ 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

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

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

$$X \sim f$$

$$\Rightarrow \bar{X}_n \xrightarrow{n \geq 30} N\left(\mu, \frac{\sigma^2}{n}\right)$$

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

Stima di π $\sim$ $X \sim \text{Bee}(\pi)$

la distribuzione esatta di $\hat{\pi}$ può essere ottenuta facendo riferimento ad una binomiale

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

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

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

Stima di $\sigma^2 \rightarrow$ lo stimatore "naturale" sarebbe:

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

questo richiederebbe
la conoscenza di μ
(poco plausibile)

STIMATORE
VARIANZA
CAMPIONARIA

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

si può stimare μ
usando $\bar{X}_n$

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

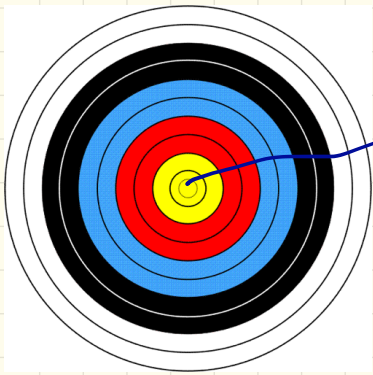
$$s^2 = \frac{\sum (X_i - \bar{X}_n)^2}{n-1}$$

STIMATORE
VARIANZA
CAMPIONARIA
CORRETTA

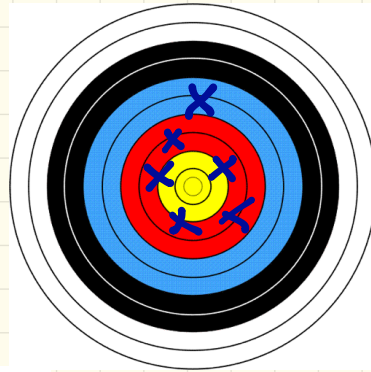
La distribuzione campionaria di $\bar{x}$ richiede l'introduzione di un nuovo modello di probabilità: la v.c. χ^2 $\rightarrow$ caratterizzata da un unico parametro, detto gradi di libertà (g.d.l.)

$\downarrow$
è definita dalle
somme di n v.c. normali
standardizzate al quadrato

$\swarrow$
nel caso dell'inferenza su varianze è necessaria l'ipotesi di normalità (rispetto ai problemi relativi alle medie o alle proporzioni, l'inferenza sulle varianze è più sensibile a violazioni dell'ipotesi di normalità)



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

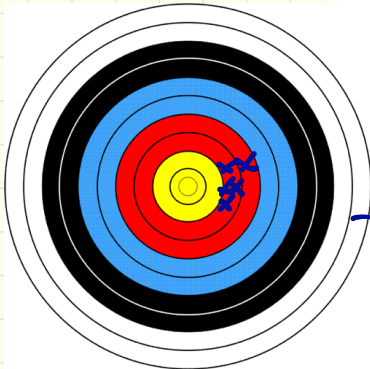


comportamento
medio



$$E(T_m) = \Theta$$

T_m è non distorto per Θ



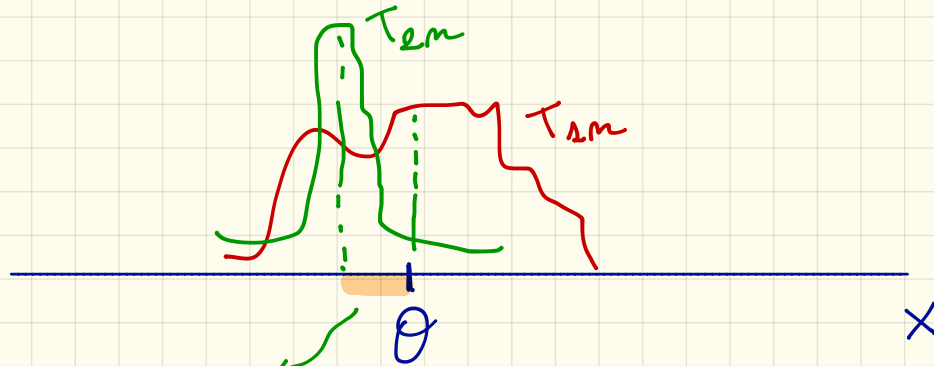
$$E(T_m) \neq \Theta$$

qui T_m è distorto: $E(T_m) - \Theta =$
 $= d(T_m) = b(T_m)$

distorsione (bias)

Se $b(T_m) < 0 \leadsto$ sottostimando θ

$b(T_m) > 0 \leadsto$ sovrastimando θ



misura
della distorsione
di T_m

In sintesi la distorsione
sintetica $E(T_m)$

↓
posizione delle v.c. stimate

$$\bar{X}_m = \frac{\sum X_i}{m}$$



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

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



le medie campionarie
 $\bar{x}$ non distorcono per $\theta = \mu$

$$J_m = \bar{X}_m \quad \text{particolare media di } \underline{\text{osservazioni } 0, 1}$$



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



estratta da una
 popolazione $X \sim \text{Ber}(\pi)$

J_m è uno stimatore non
 distorto per π

$J_m = \hat{\pi}$ (altro possibile
 nome)

Immaginiamo di essere interessati a stimare σ^2
↓
il nostro σ

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 reale 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

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

(property variance)

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

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

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

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

$$\Downarrow$$

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

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

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

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

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

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

$$\text{Var}(X_i) = \text{Var}(X)$$

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

← la distorsione diminuisce al crescere di n

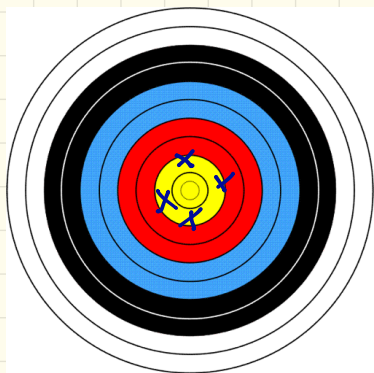
$$n \rightarrow \infty \text{ allora } \frac{n-1}{n} \rightarrow 0$$

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

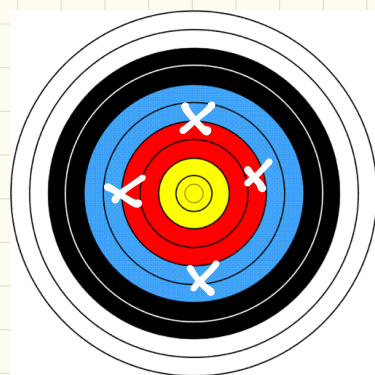
$$\begin{aligned} E(S^2) &= E\left(\frac{n}{n-1} \hat{\sigma}^2\right) = \frac{n}{n-1} E(\hat{\sigma}^2) = \\ &= \frac{n}{n-1} \frac{n-1}{n} \sigma^2 = \sigma^2 \end{aligned}$$



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



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



$$\begin{aligned} \text{Var}(T_{1m}) &= E[T_{1m} - E(T_{1m})]^2 = \\ &= E[T_{1m} - \theta]^2 \end{aligned}$$

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

$$\text{Var}(T_{1m})$$

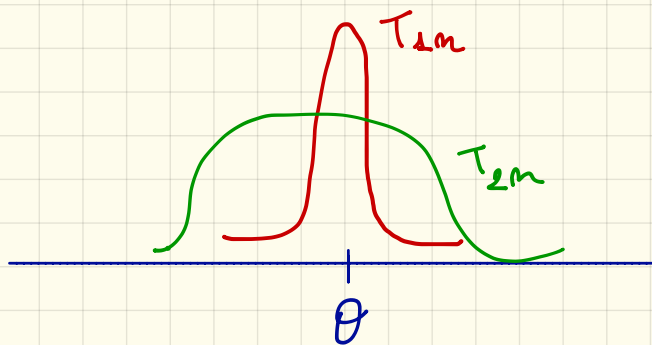
$$\text{Var}(T_{2m})$$

T_{1m} è preferibile



sono equivalenti

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



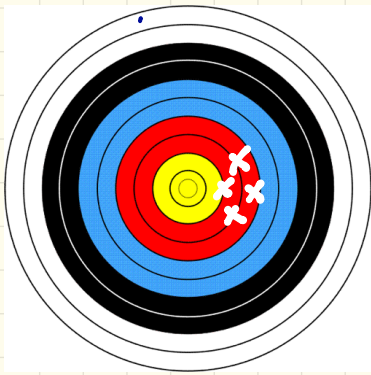
Efficienza in senso relativo consiste in un confronto tra le misure di variabilità delle distribuzioni empiriche dei due stimatori

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

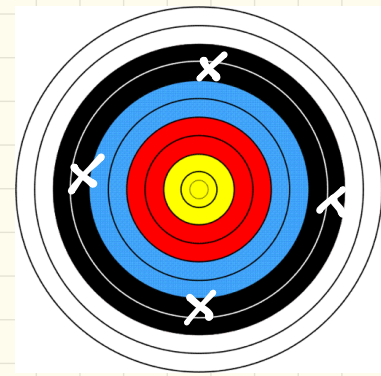
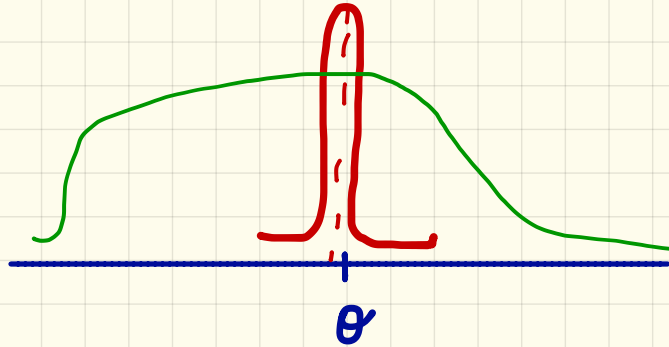
limite di
Cramér - Rao



Se $\text{var}(T_m) = \text{valore limite}$ } T_m è efficiente in senso assoluto



T_{sm}



T_m

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

che tiene conto contemporaneamente delle due proprietà

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

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

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

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

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

||
0

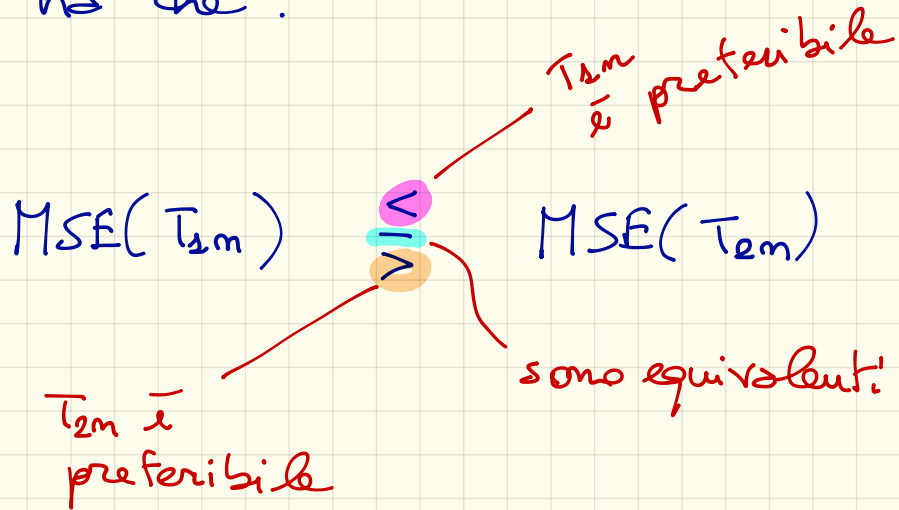
variabile
scarto

↓
E(costante) = 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

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

popolazione

obiettivo stima

simmetrica

$$\mu = M_e$$

$\bar{X}_m$ media campionaria

$$\hat{M}_e$$

media campionaria

non distorti $\leadsto E(\bar{X}_m) = E(\hat{M}_e) = \mu$

$$\text{Var}(\bar{X}_m) = \frac{\sigma^2}{n} < \frac{\pi \sigma^2}{2n} = \text{Var}(\hat{M}_e)$$

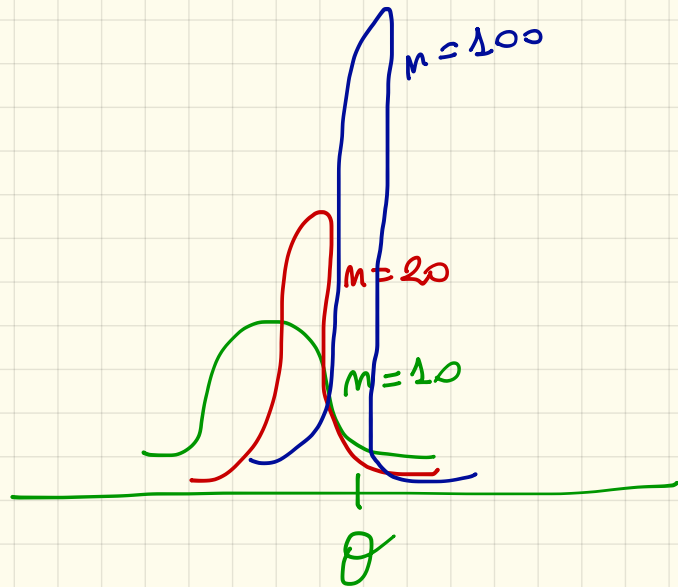
$$\frac{\text{Var}(\hat{M}_e)}{\text{Var}(\bar{X}_m)} = \frac{\pi \sigma^2 / 2n}{\sigma^2 / n} = \frac{\pi}{2} \approx 1.57$$

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

$$E(\hat{\sigma}^2) \neq \sigma^2 \quad \text{distorto} \quad \longrightarrow \quad \text{distorto non asintoticamente}$$

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

$$\hookrightarrow \lim [E(T_n - \theta)^2] = 0$$



La consistenza in
medie quadratiche vale
se e solo se:

$$① \quad \lim_{n \rightarrow \infty} E(T_n) = \theta$$

T_n è asintoticamente
non distorto

$$② \quad \lim_{n \rightarrow \infty} \text{Var}(T_n) = 0$$

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

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

RIEPILOGO DISTRIBUZIONI CAMPIONARIE

CASO DI INFERENZA SU UNA SOLA POPOLAZIONE

Schema sintesi: (lessico di base)

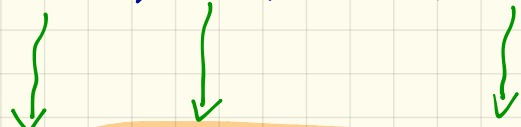
$$X \sim f(x; \theta)$$

modello
popolazione

parametro o vettore di
parametri

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

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 sintesi campionarie su tutti i "possibili" campioni $\rightarrow$ la scelta della statistica (stimatore) da usare si basa sulle caratteristiche della corrispondente distribuzione campionaria

$$\begin{array}{ccc} T(X_1, X_2, \dots, X_n) = T_n & \text{stimatore (v.c.)} \\ \downarrow \quad \downarrow \quad \downarrow \quad \downarrow & \\ T(x_1, x_2, \dots, x_n) = t_n & \text{stima} \end{array}$$

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



$\bar{X}_n$ $\left\{ \begin{array}{l} E(\bar{X}_n) = \mu \\ \text{Var}(\bar{X}_n) = \frac{\sigma^2}{n} \end{array} \right. \text{ consistente}$

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

② come si distribuisce ②

proprietà riproduttiva $\rightarrow$ 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)

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

σ

incognito [PARAMETRO DI DISTURBO]

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

s

uso il campione per stimare anche il parametro di disturbo

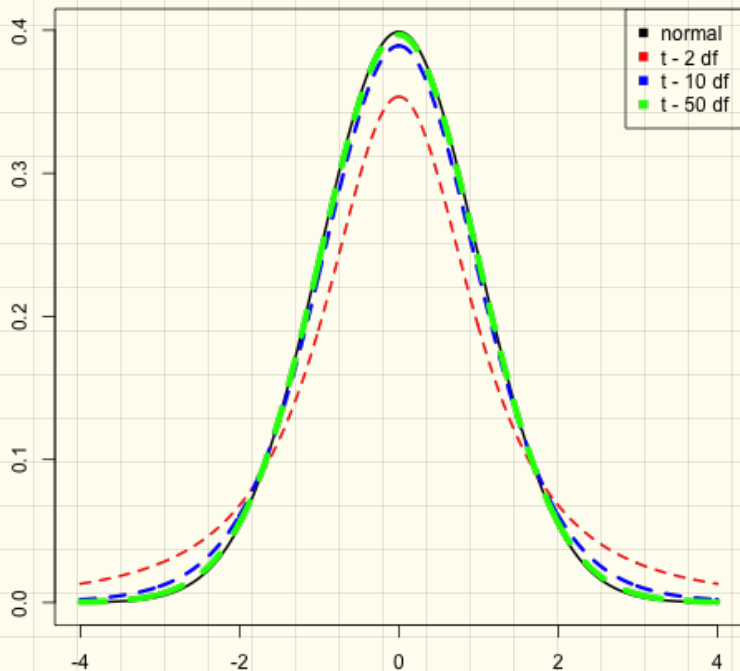
t_{n-1}

$\downarrow$

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 tabelle 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 stimare μ sulla base delle osservazioni campionarie

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

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

se n cresce

INFÈRENZA SULLA VARIANZA σ^2

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

di solito non
ho informazioni
su μ : parametro di
disturbo

σ

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
violazione dell'ipotesi di normalità

INFERENZA SULLA PROPORZIONE π

$$X \sim \text{Ber}(\pi)$$

↳ in questo caso la distribuzione esatta di $\bar{X}_n = \bar{X}_n = \hat{\pi}$ 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

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

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

$$X \sim N(\mu_x, \sigma_x^2)$$

↓
nota a priori

$$Y \sim N(\mu_y, \sigma_y^2)$$

↓
nota a priori

⇓

$$\bar{X}_m \sim N(\mu_x, \frac{\sigma_x^2}{m_x})$$

per le proprietà riproduttive della normale

$$\bar{Y}_m \sim N(\mu_y, \frac{\sigma_y^2}{m_y})$$

⇓

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

$$\bar{X}_m - \bar{Y}_m \sim N(\mu_x - \mu_y, \frac{\sigma_x^2}{m} + \frac{\sigma_y^2}{m})$$

↓
qualsiasi combinazione lineare di v.c. normali è ancora normale

$$\frac{(\bar{X}_m - \bar{Y}_m) - (\mu_x - \mu_y)}{\sqrt{\frac{\sigma_x^2}{m} + \frac{\sigma_y^2}{m}}} \sim N(0, 1)$$

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 della 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 OMOSCHEDASTICITA': le due popolazioni hanno varianza incognita che però si può assumere uguale $\rightarrow \sigma_x^2 = \sigma_y^2 = \sigma^2$

Si ha allora:

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



$$\bar{X}_m \sim N(\mu_x, \frac{\sigma^2}{m_x})$$

$$Y \sim N(\mu_y, \sigma^2)$$



$$\bar{Y}_m \sim N(\mu_y, \frac{\sigma^2}{m_y})$$

$$\frac{(\bar{X}_m - \bar{Y}_m) - (\mu_x - \mu_y)}{\sqrt{\frac{\sigma^2}{m_x} + \frac{\sigma^2}{m_y}}} = \frac{(\bar{X}_m - \bar{Y}_m) - (\mu_x - \mu_y)}{\sqrt{\sigma^2 \left(\frac{1}{m_x} + \frac{1}{m_y} \right)}}$$

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

posso stimarlo usando lo stimatore congiunto s_p^2

$$\rightarrow \frac{(\bar{X}_m - \bar{Y}_m) - (\mu_x - \mu_y)}{\sqrt{s_p^2 \left(\frac{1}{m_x} + \frac{1}{m_y} \right)}} \sim t_{m_x + m_y - 2}$$

p sta per pooled (congiunto)

$$S_p^e = \frac{(m_x - 1) S_x^e + (m_y - 1) S_y^e}{m_x + m_y - 2}$$

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^e e S_y^e , si ottiene:

$$S_p^e = \frac{\frac{(m_x - 1) \sum_i (x_i - \bar{x}_{m_x})^e}{(m_x - 1)} + \frac{(m_y - 1) \sum_i (y_i - \bar{y}_{m_y})^e}{(m_y - 1)}}{(m_x + m_y - 2)} =$$

$$= \frac{\sum_i (x_i - \bar{x}_{m_x})^e + \sum_i (y_i - \bar{y}_{m_y})^e}{m_x + m_y - 2} = \frac{\text{Dev}(X) + \text{Dev}(Y)}{m_x + m_y - 2}$$

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

$$X \sim f$$

rimuoviamo l'ipotesi di normalità delle due popolazioni

$$Y \sim f$$

$$\bar{X}_n \not\sim N\left(\mu_x, \frac{\sigma_x^2}{n_x}\right)$$

$\rightarrow$
($n_x \geq 30$)

$$\bar{Y}_n \not\sim N\left(\mu_y, \frac{\sigma_y^2}{n_y}\right)$$

$\rightarrow$
($n_y \geq 30$)

$$\bar{X}_n - \bar{Y}_n \not\sim N\left(\mu_x - \mu_y, \frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}\right)$$

$\rightarrow N$

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 faremo sotto l'ipotesi di omoschedasticità

CONFRONTO TRA LE VARIANZE DI DUE POPOLAZIONI NORMALI

$$X \sim N(\mu_x, \sigma_x^2)$$



$$\frac{(n_x - 1) s_x^2}{\sigma_x^2} \sim \chi_{n_x - 1}^2$$

$$Y \sim N(\mu_y, \sigma_y^2)$$



$$\frac{(n_y - 1) s_y^2}{\sigma_y^2} \sim \chi_{n_y - 1}^2$$

$$\frac{\cancel{(n_x - 1)} s_x^2 / \cancel{\sigma_x^2} \cancel{(n_x - 1)}}{\cancel{(n_y - 1)} s_y^2 / \cancel{\sigma_y^2} \cancel{(n_y - 1)}}$$

||

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

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

$$\frac{s_x^2 / \sigma_x^2}{s_y^2 / \sigma_y^2} \sim F_{\underbrace{n_x - 1}_{\text{gdl del numeratore}}, \underbrace{n_y - 1}_{\text{gdl del denominatore}}}$$

due parametri

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

$$X \sim \text{Ber}(\pi_x)$$



la distribuzione esatta di $F_x = \bar{X}_n = \hat{\pi}_x$
è ricavabile a partire dalla v.c. binomiale



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

$$\hookrightarrow (\hat{\pi}_x - \hat{\pi}_y) \xrightarrow{\text{TLC}} N \left(\pi_x - \pi_y, \frac{\pi_x(1-\pi_x)}{n_x} + \frac{\pi_y(1-\pi_y)}{n_y} \right)$$



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

$$\hat{\pi}_p = \frac{\sum X_i + \sum Y_i}{n_x + n_y}$$

proporzione di successi nei
due campioni

$$Y \sim \text{Ber}(\pi_y)$$



lo stesso per
 $F_y = \bar{Y}_n = \hat{\pi}_y$



$$\frac{(\hat{\pi}_x - \hat{\pi}_y) - (\pi_x - \pi_y)}{\sqrt{\hat{\pi}_p (1 - \hat{\pi}_p) \left(\frac{1}{m_x} + \frac{1}{m_y} \right)}} \xrightarrow{ZLC} N(0, 1)$$

dove: $\hat{\pi}_p = \frac{\sum_{i=1}^{m_x} X_i + \sum_{i=1}^{m_y} Y_i}{m_x + m_y}$