Università degli Studi di Napoli “Parthenope”

Corso di Laurea in Statistica e Informatica per l'Azienda, la Finanza e le Assicurazioni (SIAFA)

STATISTICA
II MODULO

Sergio LONGOBARDI
longobardi@uniparthenope.it

Inferenza

L'inferenza statistica utilizza le informazioni raccolte su un campione per conoscere parametri incogniti della popolazione

Inferenza ⇨ Processo induttivo di stima di un parametro,soggetto a incertezza perché basato su informazione parziale (quella contenuta nel campione)

La stima prodotta è soggetta ad errore (errore campionario)

stima = parametro ± errore

Inferenza

Parametro

costante non nota della popolazione,

grandezza caratteristica oggetto di inferenza (media, varianza e proporzione della popolazione)

Statistica

funzione delle osservazioni campionarie utilizzata per stimare il parametro incognito (media, varianza e proporzione campionarie)

Campionamento

- ***Una Popolazione finita*** è un insieme di N unità su cui si può osservare un certo carattere. (es: gli investimenti annui di tutte le aziende di un paese; il numero di figli di ogni famiglia italiana)
- I ***parametri*** della popolazione sono delle costanti che descrivono aspetti caratteristici della distribuzione del carattere nella popolazione stessa.

Esempio:

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

Campionamento

- Una **Popolazione infinita** è composta da tutte le unità *potenzialmente osservabili* e non necessariamente già esistenti fisicamente.

- Il carattere d'interesse può essere rappresentato da una variabile casuale con una certa distribuzione di probabilità. In questo caso si indicherà con "*popolazione X*" la variabile casuale X.

Esempio:

Media della popolazione: $\mu = E(X)$

Varianza della popolazione: $\sigma^2 = E[X - E(X)]^2$

Campionamento

I campioni possono essere estratti casualmente dalla popolazione:

• con ripetizione: una volta estratta un’unità viene rimessa dentro la popolazione e quindi potrebbe essere nuovamente estratta;

• senza ripetizione: una volta estratta un’unità questa viene messa da parte e quindi non può essere estratta più di una volta.

Due campioni **non ordinati** di uguale numerosità sono diversi tra loro se almeno un’unità del primo campione non è contenuta nel secondo campione.
Nei campioni **ordinati** conta invece anche l’ordine con cui si presentano le diverse unità.

Campionamento

Popolazione composta da 4 grandi aziende. (N=4); Carattere="Fatturato annuo";

$$x_1 = 52, x_2 = 49, x_3 = 65, x_4 = 74$$

Spazio campionario Ω, costituito dai campioni *ordinati* di dimensione 3, estratti *senza ripetizione*.

$c_1 =$	52 49 65	$c_7 =$	49 65 74	$c_{13} =$	65 74 52	$c_{19} =$	74 52 49
$c_2 =$	52 65 49	$c_8 =$	49 74 65	$c_{14} =$	65 52 74	$c_{20} =$	74 49 52
$c_3 =$	49 52 65	$c_9 =$	65 49 74	$c_{15} =$	74 52 65	$c_{21} =$	52 74 49
$c_4 =$	49 65 52	$c_{10} =$	65 74 49	$c_{16} =$	74 65 52	$c_{22} =$	52 49 74
$c_5 =$	65 52 49	$c_{11} =$	74 49 65	$c_{17} =$	52 65 74	$c_{23} =$	49 52 74

Ogni campione ha uguale probabilità di essere estratto, pari a 1/24

Campionamento

Nel campionamento casuale semplice i campioni di uguale dimensione hanno tutti stessa probabilità di essere estratti.

- **si devono conoscere le unità della popolazione;**
- **tutte le unità devono essere reperibili;**
- **si deve procedere all'estrazione casuale delle unità.**

estrazione → con ripetizione

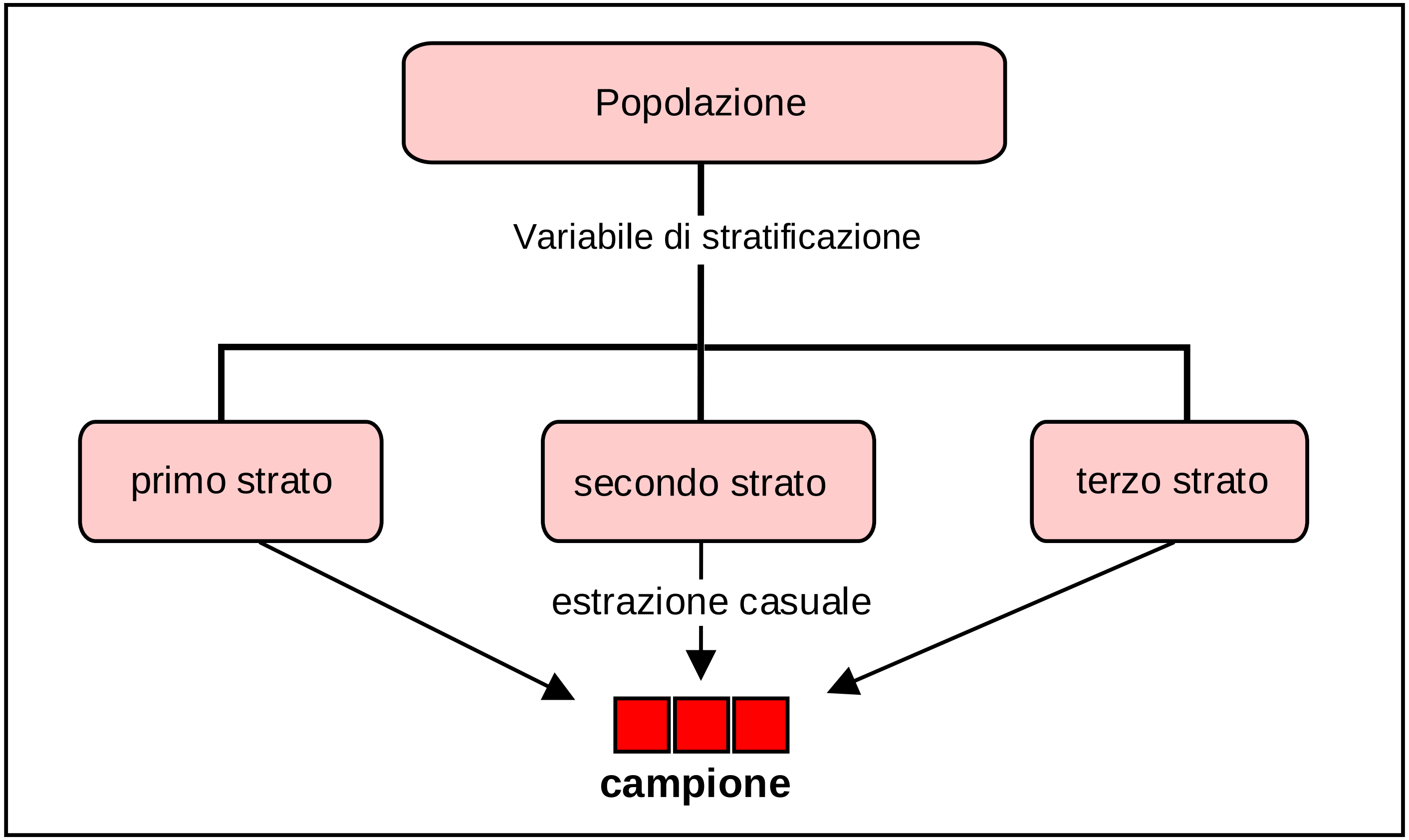
estrazione → senza ripetizione

Il campione casuale ottenuto con *estrazioni senza ripetizione* è composto da *n* variabili casuali che hanno marginalmente la stessa distribuzione di probabilità ma non sono indipendenti.

La distribuzione di probabilità della generica X_i è uguale a quella del carattere *X* nella popolazione.

Campionamento

Nel campionamento casuale stratificato la popolazione viene suddivisa in strati. Da ogni strato vengono poi estratti, tramite un campionamento casuale semplice, le unità da inserire nel campione. *Esempio strati: Regioni; età; sesso*.



Campionamento casuale a grappoli e a stadi
Popolazione Finita

Nel campionamento casuale a grappoli la popolazione viene suddivisa in sottoinsiemi detti grappoli. Si selezionano, con un'estrazione casuale senza ripetizione, un certo numero di grappoli e si prendono come unità campionarie tutte le unità appartenenti ai grappoli estratti.

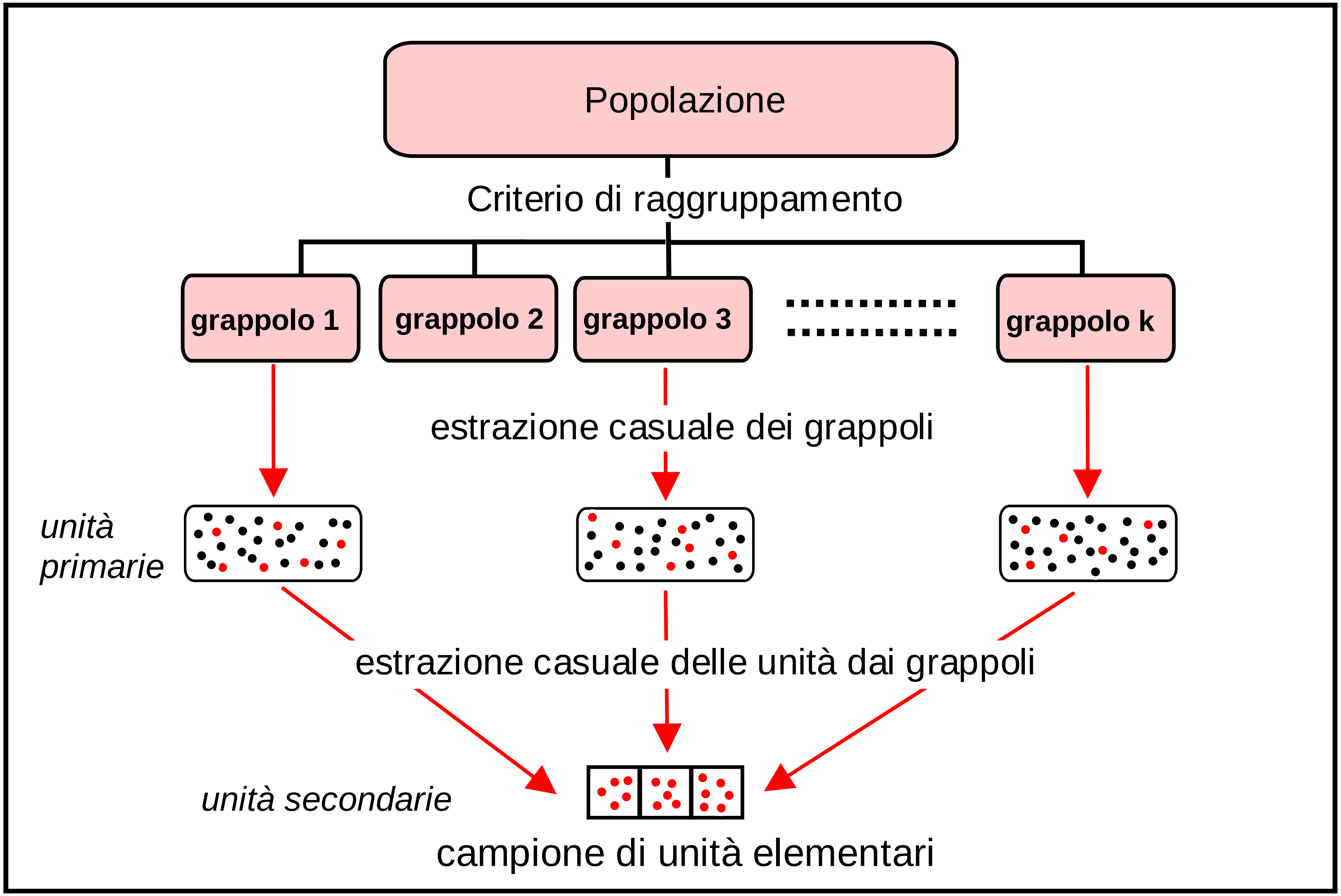
Nel campionamento casuale a due stadi la popolazione viene suddivisa in un certo numero di grappoli. Al primo stadio si estrae senza ripetizione un certo numero di grappoli. Da ciascuno di questi si estrae con ripetizione (secondo stadio) un certo numero di unità.

primo stadio

secondo stadio

Unità primarie → **Unità secondarie**

Campionamento casuale a grappoli e a stadi
Popolazione Finita



Campionamento da popolazioni infinite

In una popolazione infinita, la *n*-pla di variabili casuali:

$$(X_1, X_2, \ldots, X_n)$$

che compongono il campione casuale di dimensione *n* presenta le seguenti proprietà:

- **$X_1, X_2, \ldots, X_n$ sono variabili casuali indipendenti.**

- **ogni v.c. X_i possiede la stessa distribuzione di probabilità della popolazione *X*.**

Nelle popolazioni finite in cui la dimensione campionaria è molto più piccola della numerosità della popolazione, si può applicare la teoria del campionamento da popolazioni infinite.

Statistiche campionarie e distribuzioni campionarie

Una statistica campionaria è una funzione a valori reali delle osservazioni campionarie:

$$T = t\,(X_1, X_2, \ldots, X_n)$$

Esempi di statistiche campionarie:

media campionaria:

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

varianza campionaria corretta:

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

La statistica campionaria è una variabile casuale a cui è associata una distribuzione di probabilità detta distribuzione campionaria.

Statistiche campionarie e distribuzioni campionarie - Esempio

Consideriamo una popolazione finita composta dalle seguenti 5 unità:

$$x_1 = 8 \quad x_2 = 4 \quad x_3 = 2 \quad x_4 = 11 \quad x_5 = 6$$

Si consideri l'estrazione senza ripetizione di campioni di dimensione *n*=2 e per ognuno di essi si calcoli la statistica *media campionaria*.

X_1	8	8	8	8	4	4	4	4	2	2	2	2	11	11	11	11	6	6	6	6
X_2	4	2	11	6	8	2	11	6	8	4	11	6	8	4	2	6	8	4	2	11
$\overline{X}$	6	5	9,5	7	6	3	7,5	5	5	3	6,5	4	9,5	7,5	6,5	8,5	7	5	4	8,5

La distribuzione di probabilità della media campionaria è data da:

$\overline{X}$	3	4	5	6	6,5	7	7,5	8,5	9,5
$P(\overline{X})$	0,1	0,1	0,2	0,1	0,1	0,1	0,1	0,1	0,1

Distribuzione della media campionaria nelle popolazioni infinite

Proprietà della media campionaria :

il valore atteso $E(\bar{X}) = \mu$

la varianza $Var(\bar{X}) = \sigma^2 / n$

Se $X \sim N(\mu; \sigma^2)$ **allora** $\bar{X} \sim N\left(\mu; \frac{\sigma^2}{n}\right)$

Qualunque sia la popolazione, per il Teorema del Limite Centrale *(vedi slides successive)*

$$\lim_{n \to \infty} P\left(\frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \leq z\right) = P(Z \leq z)$$

dove Z è una v.c. Normale standardizzata

Distribuzione della media campionaria nelle popolazioni finite

Si consideri una popolazione finita dalla quale viene estratto senza ripetizione un campione casuale. In questo caso:

- **il valore atteso**

$$E(\overline{X}) = \mu$$

- **la varianza**

$$Var(\overline{X}) = \left(\frac{N-n}{N-1}\right)\frac{\sigma^2}{n}$$

- **Se *n* è sufficientemente ampio ma molto più piccolo di *N*, allora la distribuzione di $\overline{X}$ può essere approssimata a una Normale con media μ e varianza**

$$\left(\frac{N-n}{N-1}\right)\frac{\sigma^2}{n}$$

Teorema limite centrale

Il teorema del limite centrale (TLC) afferma che la somma (o la media) di un numero elevato di variabili aleatorie indipendenti e identicamente distribuite (i.i.d.) è approssimativamente normale **indipendentemente** dalla distribuzione soggiacente.

Teorema limite centrale

Il TLC implica che se la grandezza di un campione è ELEVATA, allora la distribuzione della somma S_n di n variabili aleatorie indipendenti (oppure la loro media) sarà approssimativamente normale.

Il TLC comporta che la distribuzione di alcune statistiche (per esempio la media del campione) diventa nota anche se non si hanno conoscenze sulla distribuzione della popolazione da cui i campioni sono stati estratti

Teorema limite centrale

Date X_1, X_2,…,X_n v.c. indipendenti e identicamente distribuite (i.i.d.) con media

μ e varianza σ^2 la v.c. somma:

$$S_n = \sum_{i=1}^{n} X_i$$

al ***crescere di n*** tende a distribuirsi come una Normale con media $n\mu$ e varianza $n\sigma^2$:

$$S_n \sim N(n\mu, n\sigma^2)$$

Se si considera la standardizzazione della v.c. somma S_n si ottiene una v.c. standardizzata Z_n (con media 0 e varianza 1) definita come:

$$Z_n = \frac{S_n - E(S_n)}{\sqrt{Var(S_n)}}$$

Teorema limite centrale

Il TLC comporta che la distribuzione di alcune statistiche (per esempio la media del campione) diventa nota anche se non si hanno conoscenze sulla distribuzione della popolazione da cui i campioni sono stati estratti

$$S_n \sim N(n\mu, n\sigma^2) \qquad Z_n = \frac{S_n - E(S_n)}{\sqrt{Var(S_n)}}$$

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{S_n}{n} \Rightarrow Z_n = \frac{\frac{\sum_{i=1}^{n} X_i}{n} - \frac{n\mu}{n}}{\frac{\sqrt{n\sigma^2}}{n}}$$

$$Z_n = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$