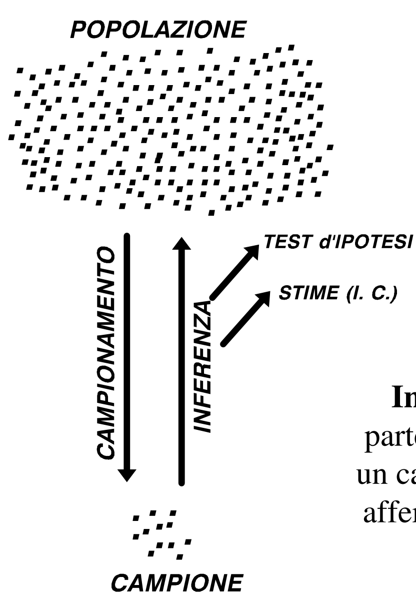
Campionamento e distribuzioni campionarie

Prof. Giuseppe Verlato
Sezione di Epidemiologia e Statistica Medica,
Università di Verona



Inferenza statistica = processo che, partendo dalle informazioni contenute in un campione, consente di giungere a delle affermazioni relative alla popolazione da cui proviene quel campione.

Il campione deve essere **rappresentativo** della popolazione originaria.

Se si intervistano i passanti in una determinata strada cittadina, si ottiene un campione rappresentativo della popolazione originaria?

No, in quanto la probabilità di essere intervistati varia da un soggetto (unità statistica) all'altro a seconda delle abitudini personali ed è ignota.

Per ottenere un campione rappresentativo, bisogna disporre di una **lista di campionamento** *(sampling frame)* da cui estrarre il campione.

Devo conoscere esattamente la probabilità che un'unità statistica ha di essere estratta (**campionamento di tipo probabilistico).**

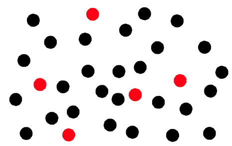
Nella notte delle elezioni europee del 2004, un politico ha contestato uno studio campionario effettuato da un ente di ricerca statistica.

Infatti la lista del politico aveva ottenuto il 31,5% dei suffragi stando alle proiezioni effettuate dall'ente di ricerca, mentre aveva ottenuto il 34,7% sulla base dei dati del Viminale, che aveva assemblato i risultati del 20% delle sezioni, scrutinate per prime.

Il politico ha sottolineato che i dati dello statistico erano soltanto delle proiezioni, mentre i dati del Viminale erano dei dati reali.

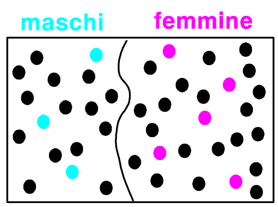
Lo statistico gli ha rispettosamente risposto che i dati del Viminale, pur coprendo il 20% dell'intera popolazione, non erano rappresentativi del territorio nazionale, in quanto i risultati di alcune regioni affluivano per primi.

Il giorno dopo i giornali hanno pubblicato che la lista del politico aveva effettivamente preso il 31,5% dei suffragi.



CAMPIONAMENTO CASUALE SEMPLICE:
tutte le unità di osservazione hanno la stessa probabilità di essere selezionate (frazione di campionamento = n/N)

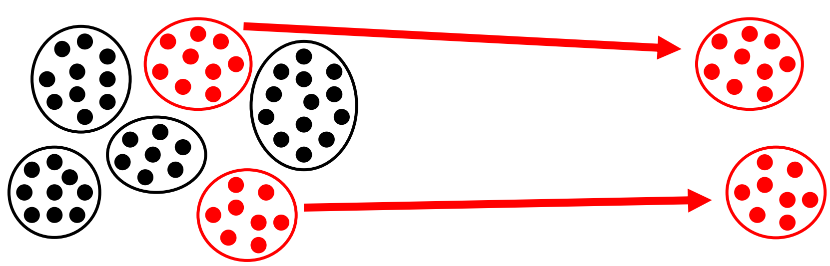
CAMPIONAMENTO SISTEMATICO: si sfrutta un ordinamento naturale delle unità di osservazione: si campiona a caso uno dei primi k soggetti e poi si mantiene un passo di campionamento k costante



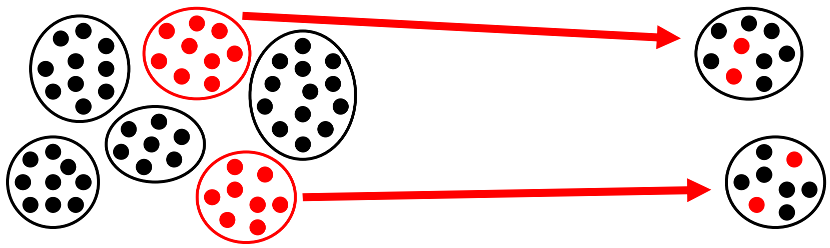
CAMPIONAMENTO STRATIFICATO:
viene effettuato un campionamento casuale semplice in ogni STRATO della popolazione. Allocazione ottimale:
n_i proporzionale a $N_i \sigma_i$

CAMPIONAMENTO A GRAPPOLO (CLUSTER):

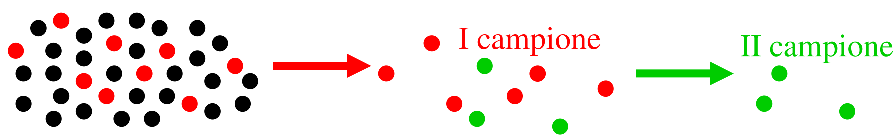
sono estratte a caso delle unità campionarie di II livello, e vengono reclutate tutte le unità campionarie di I livello in esse contenute.



CAMPIONAMENTO A DUE O PIU' STADI: sono estratte a caso delle **unità campionarie di II livello**, all'interno delle quali vengono estratte a caso delle **unità campionarie di I livello**.



CAMPIONAMENTO A DUE O PIU' FASI:
viene estratto un primo campione da sottoporre a indagini meno approfondite, dal quale viene estratto un secondo campione da sottoporre ad indagini più approfondite.



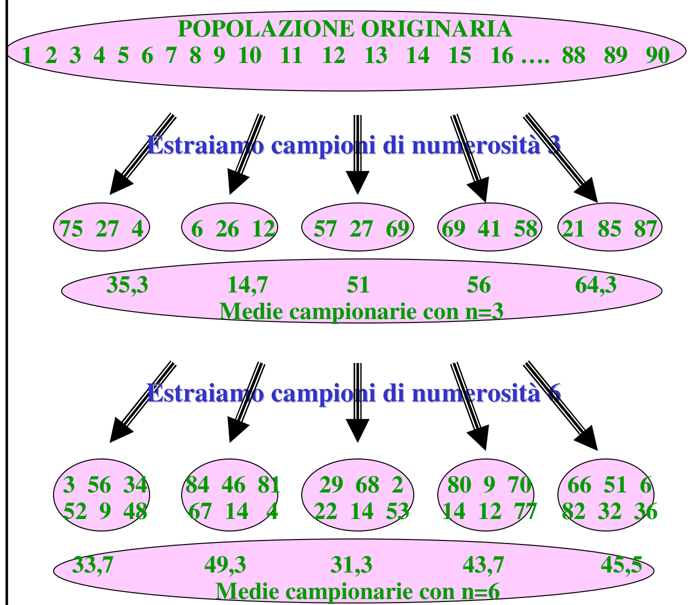
La media campionaria è una **variabile casuale**.

Infatti:

1) varia da un campione all'altro

2) i campioni sono estratti in modo casuale.

DETERMINAZIONE SPERIMENTALE della DISTRIBUZIONE di una MEDIA CAMPIONARIA - 1



$\Sigma x=4095$ $\mu=45,5$
$\Sigma x^2=247065$ $\sigma=26,12$
distribuzione uniforme

$\Sigma x=221,3$ $\bar{x}_x=44,3$
$\Sigma x^2=11339,3$ $s_x=19,6$

$\Sigma x=203,5$ $\bar{x}_x=40,7$
$\Sigma x^2=8526$ $s_x=7,8$

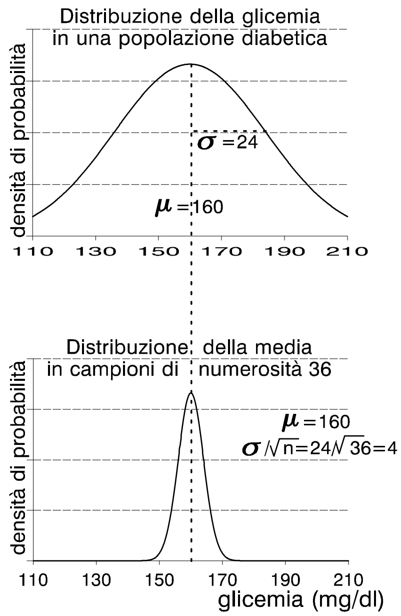
DETERMINAZIONE SPERIMENTALE della DISTRIBUZIONE di una MEDIA CAMPIONARIA - 2

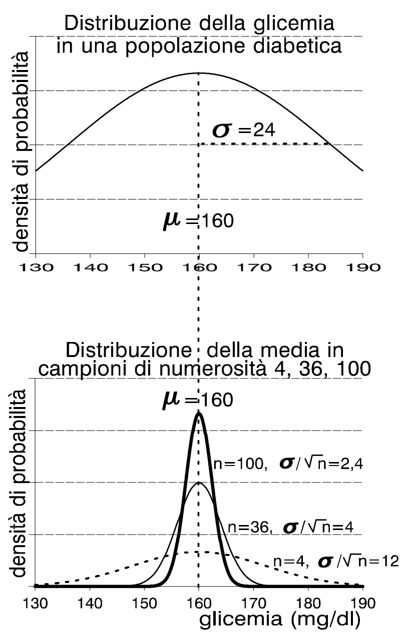
1) La **media delle medie** (**44,27 per n=3** e **40,7 per n=6**) è all'incirca **uguale** alla **media della popolazione** originaria (**45,5**)

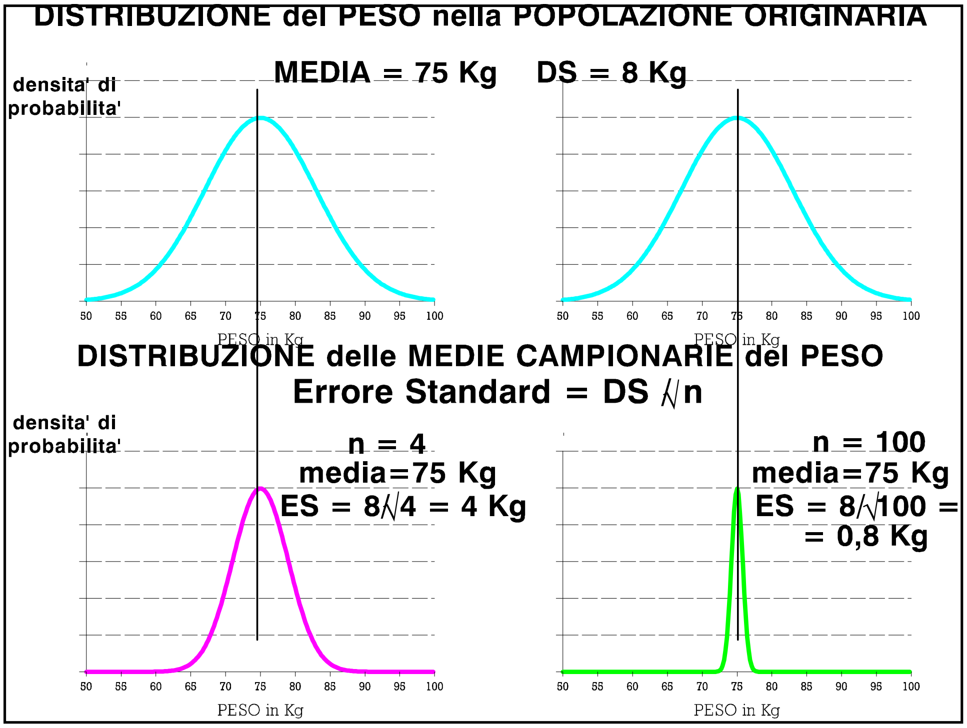
2) La **variabilità delle medie campionarie** è **inferiore** alla **variabilità osservata nella popolazione** originaria: DS_x=Errore Standard = $\sigma/\sqrt{n}$

numerosità campionaria	valore osservato	valore atteso
n = 3	$DS_x = 19,63$	$26,12/\sqrt{3}=15,08$
n = 6	$DS_x = 7,80$	$26,12/\sqrt{6}=10,67$

3) **All'aumentare della numerosità campionaria**, **la media campionaria tende a distribuirsi normalmente** indipendentemente dalla distribuzione della variabile originaria

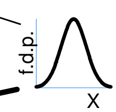

Distribuzione della glicemia
in una popolazione diabetica
densità di probabilità
σ =24
μ =160
110 130 150 170 190 210
Distribuzione della media
in campioni di numerosità 36
μ =160
σ /√n=24/√36=4
110 130 150 170 190 210
glicemia (mg/dl)


Distribuzione della glicemia
in una popolazione diabetica
densità di probabilità
σ =24
μ =160
130 140 150 160 170 180 190
Distribuzione della media in
campioni di numerosità 4, 36, 100
μ =160
n=100, σ /√n=2,4
n=36, σ /√n=4
n=4, σ /√n=12
130 140 150 160 170 180 190
glicemia (mg/dl)



DISTRIBUZIONE delle STATISTICHE CAMPIONARIE

STATISTICA	E(X)	E(X-E(X))^2	DISTRIBUZIONE
media ($\bar{X}$)	μ	σ^2/n	normale
mediana		$1,57\sigma^2/n$	
varianza (S^2) corretta	σ^2	$\frac{2\sigma^4}{n-1}$	χ^2_{n-1}
frequenza relativa (p)	π	$\frac{\pi(1-\pi)}{n}$	binomiale
numero di successi (np)	$n\pi$	$n\pi(1-\pi)$	binomiale



DISTRIBUZIONE delle STATISTICHE CAMPIONARIE

	STATISTICA	E(X)	$E(X-E(X))^2$	DISTRIBUZIONE
campioni indipendenti	$\bar{X}_1 - \bar{X}_2$	$\mu_1 - \mu_2$	$\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$	se $n_1, n_2 >= 30$
	$\bar{X}_1 + \bar{X}_2$	$\mu_1 + \mu_2$		
campioni dipendenti	$\bar{X}_1 - \bar{X}_2$	$\mu_1 - \mu_2$	$\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} - 2Cov(X_1, X_2)$	normale
	$\bar{X}_1 + \bar{X}_2$	$\mu_1 + \mu_2$	$\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} + 2Cov(X_1, X_2)$