STATISTICA A – K (63 ore)

Marco Riani

mriani@unipr.it

http://www.riani.it

Inferenza statistica

- Dal campione alla popolazione
- Con quale precisione si possono descrivere le caratteristiche di una popolazione sulla base delle informazioni desunte dal campione?
- Estensione avviene in termini probabilistici

Campione

- **Campione statistico = sottoinsieme di n elementi tratti da un universo statistico (N elementi)**

- **Vantaggi: risparmio di tempo e di costi**
- **Svantaggi: risultati approssimati (incertezza $\Rightarrow$ probabilità)**

Campionamento da universo finito

Esperimento aleatorio $\Rightarrow$ estrazione di un campione di numerosità n

Modo di estrazione degli elementi

- **Con reimmissione (Bernoulliana)**

1^a	2^a	…	n^a
U N	**U N**		**U N**

Prima dell'estrazione: $\Rightarrow$ n variabili aleatorie indipendenti e identicamente distribuite

Campionamento da universo finito

- **Senza reimmissione**

1^a 2^a … n^a

Le n v.a. non sono più indipendenti e identicamante distribuite

Criterio in base al quale si distinguono due campioni:

- **differiscono almeno per l'ordine**

(a, b, c)   (a, c, b)

- **differiscono almeno per un elemento**

(a, b, c)   (a, b, d)

SPAZIO DEI CAMPIONI = Ω = insieme di tutti i possibili campioni di n elementi estraibili dall'universo

Esempio: U = (a, b, c, d) N = 4 n = 2

- Estrazione con reimmissione (o bernoulliana)
  Quanti sono i possibili campioni di numerosità n=2 se l'ordine conta?

(a,a)	(a,b)	(a,c)	(a,d)
(b,a)	(b,b)	(b,c)	(b,d)
(c,a)	(c,b)	(c,c)	(c,d)
(d,a)	(d,b)	(d,c)	(d,d)

Disposizioni con rip. = $(N)^n = 4^2 = 16$

Esempio: U = (a, b, c, d) N = 4 n = 2

- Estrazione senza reimmissione (in blocco) **due campioni sono diversi solo se differiscono per almeno un elemento**
- Quanti sono i possibili campioni di numerosità n=2 se l'ordine non conta?

(a,b)	(a,c)	(a,d)
	(b,c)	(b,d)
		(c,d)

Combinazioni $C_{N,n}$

$$\binom{N}{n} = \frac{N!}{n!(N-n)!} = \frac{4!}{2!(4-2)!} = 6$$

Le distribuzioni campionarie

La v.a. media campionaria ($\overline{X}$)

Esempio

X = peso in grammi N = 4 (oggetti)

U = (20, 20, 24, 28)

μ= 23 VAR = σ^2 = 11

Distribuzione di X nell'universo

x_i	n_i	f_i
20	2	0,5
24	1	0,25
28	1	0,25
	4	1

U = (20, 20, 24, 28) spazio dei campioni (estrazione bernoulliana)

- n =2

(20,20)	(20,20)	(20,24)	(20,28)
(20,20)	(20,20)	(20,24)	(20,28)
(24,20)	(24,20)	(24,24)	(24,28)
(28,20)	(28,20)	(28,24)	(28,28)

U = (20, 20, 24, 28)

Distribuzione di X= fenomeno nell'universo

x_i	n_i	f_i
20	2	0,5
24	1	0,25
28	1	0,25
	4	1

V. A. I elemento del campione X_1 (prima dell'estrazione)

x_i	n_i	p_i
20	2	0,5
24	1	0,25
28	1	0,25
	4	1

V. A. II elemento del campione X_2

x_i	n_i	p_i
20	2	0,5
24	1	0,25
28	1	0,25
	4	1

Conclusione

$$X_1; X_2; X_3; \dots; X_n$$

Un campione prima dell’estrazione può essere considerato come una successione di v.a. indipendenti, identicamente distribuite e con distribuzione uguale a quella del fenomeno nell’universo

(20,20)	(20,20)	(20,24)	(20,28)
20	20	22	24
(20,20)	(20,20)	(20,24)	(20,28)
20	20	22	24
(24,20)	(24,20)	(24,24)	(24,28)
22	22	24	26
(28,20)	(28,20)	(28,24)	(28,28)
24	24	26	28

U = (20, 20, 24, 28)

E(U)=μ=23 VAR(U)=σ^2=11

DISTRIBUZIONE DELLA V.A. MEDIA CAMPIONARIA

$\bar{x}_i$	n_i	p_i	$\bar{x}_i \cdot p_i$	$(\bar{x}_i-\mu)^2 \cdot p_i$
20	4	0,25	5	2,25
22	4	0,25	5,5	0,25
24	5	0,3125	7,5	0,3125
26	2	0,125	3,25	1,125
28	1	0,0625	1,75	1,5625
	16	1	23	5,5

$E(\bar{X}) = 23 = \mu$
(media dell'universo)

$VAR(\bar{X}) = 5{,}5 = \dfrac{11}{2} = \dfrac{\sigma^2}{n}$

Caratteristiche della v.a. media campionaria da un universo (μ σ^2)

$$\overline{X} = \frac{1}{n}(X_1 + X_2 + X_3 + \cdots + X_n)$$

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

$$VAR(\overline{X}) = ?$$

Caratteristiche della v.a. media campionaria da un universo (μ σ²)

$$\bar{X} = \frac{1}{n}(X_1 + X_2 + X_3 + \cdots + X_n)$$

$$E(\bar{X}) = ?$$

$$E(\bar{X}) = \frac{1}{n}E(X_1 + X_2 + X_3 + \cdots + X_n)$$

$$E(\bar{X}) = \frac{1}{n}[E(X_1) + E(X_2) + E(X_3) + \cdots + E(X_n)]$$

$$E(\bar{X}) = \frac{1}{n}[\mu + \mu + \mu + \cdots + \mu] = \frac{1}{n}n\mu = \mu$$

Caratteristiche della v.a. media campionaria da un universo (μ σ²)

$$\bar{X} = \frac{1}{n}(X_1 + X_2 + X_3 + \cdots + X_n)$$

$$VAR(\bar{X}) = ?$$

$$VAR(\bar{X}) = \frac{1}{n^2} VAR(X_1 + X_2 + X_3 + \cdots + X_n)$$

$$VAR(\bar{X}) = \frac{1}{n^2}[VAR(X_1) + VAR(X_2) + VAR(X_3) + \cdots + VAR(X_n)]$$

$$VAR(\bar{X}) = \frac{1}{n^2}[\sigma^2 + \sigma^2 + \sigma^2 + \cdots + \sigma^2] = \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n}$$

Caratteristiche della v.a. media campionaria

- Premessa: qualsiasi combinazione di v.c. normale è distribuita in maniera normale
- Se Universo~N(μ σ^2)

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

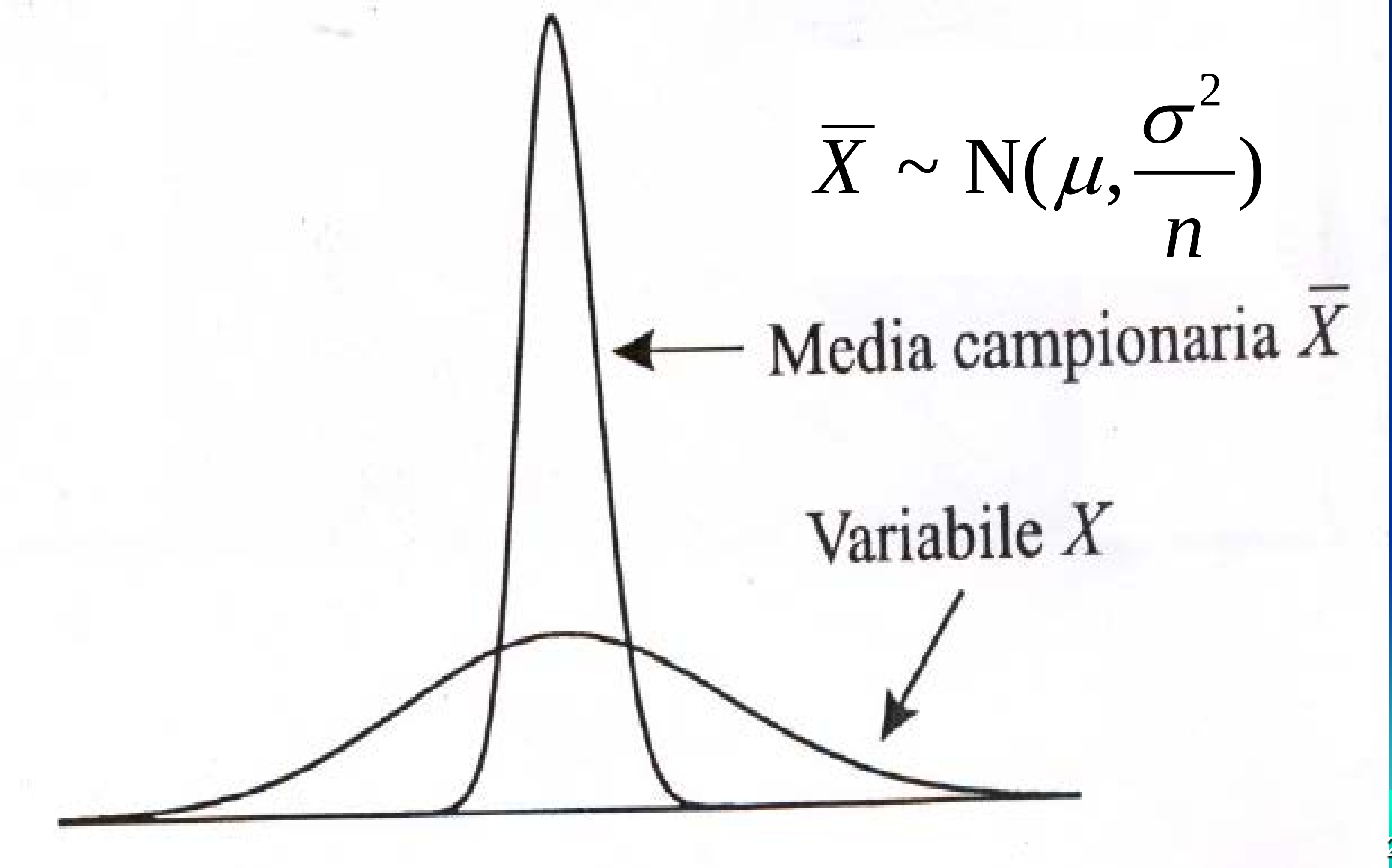
Errore standard della v.a. media campionaria

$$\bar{X} \sim \mathrm{N}(\mu, \frac{\sigma^2}{n})$$

$$\sigma(\bar{X}) = \frac{\sigma}{\sqrt{n}}$$

L'errore standard nella media quantifica la variazione della media campionaria da campione a campione

X= fenomeno nell'universo~N(μ σ^2)



Obiettivo

- Analizzare la distribuzione della v.a. media campionaria quando l’universo da cui gli elementi campionari provengono presenta forma ignota

Convergenza in legge

- Siano X, X_1 X_2 … X_n v.a., indichiamo con F, F_1 F_2 … F_n le rispettive funzioni di ripartizione. Diremo che X_n converge in legge a X se e solo se per ogni punto di continuità di F

$$\lim_{n\to\infty} F_n(x) = F(x)$$

- In simboli

$$X_n \xrightarrow{Legge} X$$

Teorema centrale del limite

Sia $X_1\ X_2 \ldots X_n$ una successione di v.a. indipendenti e identicamente distribuite con media μ e varianza σ^2.

Sia S_n la somma di queste variabili aleatorie $S_n = X_1 + X_2 + \ldots + X_n$

$E(S_n) = E(X_1 + X_2 + \ldots + X_n) = E(X_1) + E(X_2) + \ldots + E(X_n) = n\,\mu$

$VAR(S_n) = VAR(X_1 + X_2 + \ldots + X_n) = VAR(X_1) + VAR(X_2) + \ldots + VAR(X_n) = n\,\sigma^2$

$$\frac{S_n - E(S_n)}{\sqrt{\mathrm{var}(S_n)}} \xrightarrow{Legge} \mathrm{N}(0,1)$$

Teorema centrale del limite e v.a. media campionaria

$$\bar{X} = \frac{1}{n}(X_1 + X_2 + X_3 + \cdots + X_n)$$

$$\frac{S_n - E(S_n)}{\sqrt{\mathrm{var}(S_n)}} \xrightarrow{Legge} \mathrm{N}(0,1)$$

$$\frac{\bar{X} - E(\bar{X})}{\sqrt{\mathrm{var}(\bar{X})}} \xrightarrow{Legge} \mathrm{N}(0,1)$$

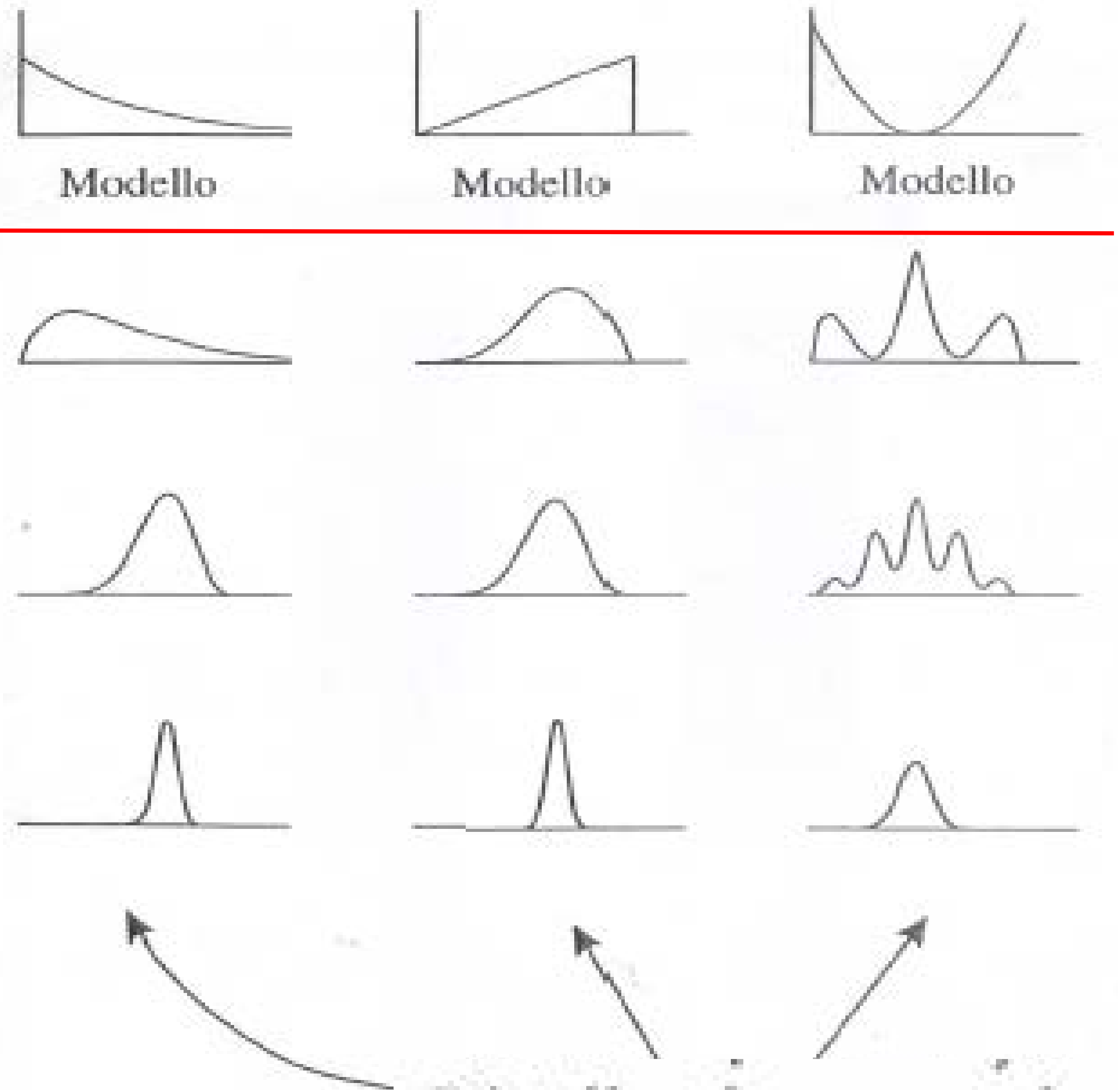
- Approssimazione sufficiente per n>=100

Distribuzione di X

n=2

n = 4

n = 25



Distribuzione campionaria di $\overline{X}$

Richiamo distribuzione della v.a. media campionaria

$$\frac{\bar{X} - E(\bar{X})}{\sqrt{\mathrm{var}(\bar{X})}} \xrightarrow{Legge} \mathrm{N}(0,1)$$

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

$$\bar{X} - \mu \sim \mathrm{N}(0, \sigma^2 / n)$$

$$\bar{X} \sim \mathrm{N}(\mu, \sigma^2 / n)$$

Caratteristiche della v.a. media campionaria da un universo (μ σ^2)

$$\bar{X} = \frac{1}{n}(X_1 + X_2 + X_3 + \cdots + X_n)$$

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

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

- Estrazione con reimmissione. Se n è elevato (grazie al teorema centrale del limite)

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

$$\bar{X} \sim \mathrm{N}(\mu, \frac{\sigma^2}{n})$$

Relazione tra v.a. Binomiale e v.a. normale

- X = Binomiale = Somma di n v.a. Bernoulliane (X_1 X_2 …, X_n) indipendenti e identicamente distribuite con Pr successo pari a π
- Teorema centrale del limite➔ se n è elevato la binomiale standardizzata ~N(0,1)

$$\frac{X - n\pi}{\sqrt{n\pi(1-\pi)}} \xrightarrow{Legge} \mathrm{N}(0,1)$$

Esercizio

- Si lancia una moneta 10000 volte . Scrivere l’espressione algebrica (ed eventualmente calcolare) la probabilità che il numero di facce testa sia compreso tra 4900 e 5100.

Esercizio

- Si lancia una moneta 10000 volte . Calcolare la probabilità che il numero di facce testa sia compreso tra 4900 e 5100.

- Universo Bernoulliano
- $n=10000$ $\qquad\qquad$ $\pi=(1-\pi)=0{,}5$
- X=numero di successi $(X_1+X_2+\ldots+X_n)$
- $X \sim B(n, \pi)=B(10000, \frac{1}{2})$

Dati del problema

- Universo Bernoulliano
- $n=10000\ \pi=(1-\pi)=0,5$
- X=numero di successi $(X_1+X_2+\dots X_n)$
- $X \sim B(n, \pi)=B(10000, ½)$

$$Pr(4900 \leq X \leq 5100) = \sum_{k=4900}^{5100} \binom{10000}{k} \left(\frac{1}{2}\right)^{k} \left(\frac{1}{2}\right)^{10000-k}$$

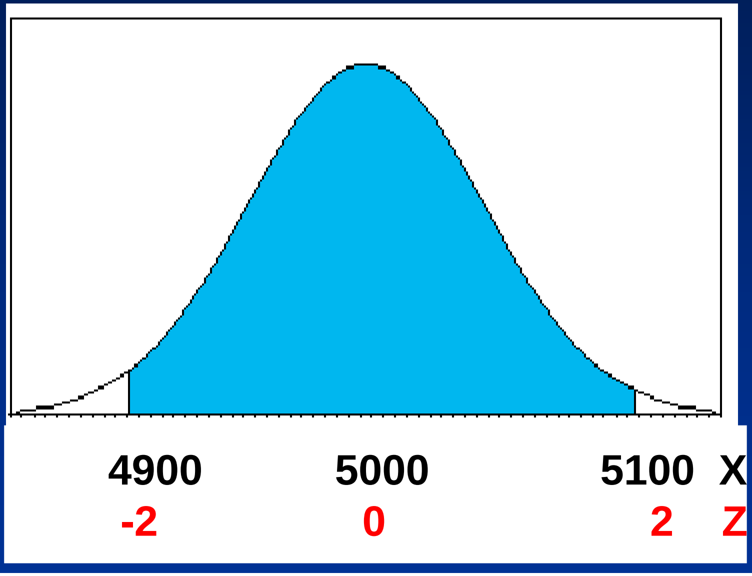
$X \sim N(\boldsymbol{n \cdot \pi}, \boldsymbol{n \cdot \pi(1-\pi)})$

$$\frac{X - n\pi}{\sqrt{n\pi(1-\pi)}} \xrightarrow{Legge} N(0,1)$$

Calcolo approssimato

X≈ N(n·π, n·π(1-π))

X≈N(5000, 50^2)



$$Pr(x_1 \leq X \leq x_2) = Pr(z_1 \leq Z \leq z_2)$$

$$Pr(4900 \leq X \leq 5100) = Pr\left(\frac{4900-5000}{50} < Z < \frac{5100-5000}{50}\right)$$

$$z_1 = \frac{4900 - 5.000}{50} = -2 \qquad z_2 = \frac{5100 - 5.000}{50} = 2$$

$F(z_2)–F(z_1)$=F(2)-F(-2)=0,97725-0,02275=0,9545

Caratteristiche della v.a. media campionaria da un universo (μ σ²)

$$\overline{X} = \frac{1}{n}(X_1 + X_2 + X_3 + \cdots + X_n)$$

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

$$VAR(\overline{X}) = \frac{\sigma^2}{n}$$

- Estrazione con reimmissione. Se n è elevato (grazie al teorema centrale del limite)

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

$$\overline{X} \sim \mathrm{N}(\mu, \frac{\sigma^2}{n})$$

Estrazione in blocco

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

$$\mathrm{var}(\overline{X}) = \frac{\sigma^2}{n} \cdot \frac{N-n}{N-1}$$

$$\sigma(\overline{X}) = \sqrt{\frac{\sigma^2}{n} \cdot \frac{N-n}{N-1}} = \frac{\sigma}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-1}}$$

- N.B. $\lim_{N\to\infty} \frac{N-n}{N-1} = 1$

$$\mathrm{var}(\overline{X}) = \frac{\sigma^2}{n} \cdot \frac{N-n}{N-1}$$

Dimostrazione

- Premessa: universo di numerosita N distribuito con ~(μ, σ^2)
- campione senza reimmissione dall'universo (X_1 X_2 … X_n)
- $X_i \sim (\mu, \sigma^2)$
- $\mathrm{Cov}(X_i, X_j)= \mathrm{cov}$ (costante)

$$\mathrm{var}(\overline{X}) = \frac{\sigma^2}{n} \cdot \frac{N-n}{N-1}$$

Dimostrazione

$$\begin{aligned} VAR(\overline{X}) &= VAR\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) \\ &= \frac{1}{n^2} VAR\left(\sum_{i=1}^{n} X_i\right) \end{aligned}$$

$$= \frac{1}{n^2}\left(\sum_{i=1}^{n} VAR(X_i) + \underbrace{\sum_{i=1}^{n}\sum_{j=1}^{n}}_{i \neq j} cov(X_i, X_j)\right)$$

$$= \frac{1}{n^2}\left(\sum_{i=1}^{n} \sigma^2 + \underbrace{\sum_{i=1}^{n}\sum_{j=1}^{n}}_{i \neq j} cov\right)$$

$$= \frac{1}{n^2}\left(n\sigma^2 + n(n-1)cov\right)$$

$$VAR(\overline{X}) = \frac{1}{n}\left(\sigma^2 + (n-1)cov\right)$$

$$\operatorname{var}(\overline{X}) = \frac{\sigma^2}{n} \cdot \frac{N-n}{N-1}$$

Dimostrazione

$$VAR(\overline{X}) \;=\; \frac{1}{n}\left(\sigma^2 + (n-1)cov\right)$$

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

$$\text{quando } n = N,\ VAR(\overline{X}) = 0$$

$$cov = -\frac{\sigma^2}{N-1}$$

Dimostrazione

$$VAR(\overline{X}) = \frac{1}{n}\left(\sigma^2 + (n-1)cov\right)$$

$$cov = -\frac{\sigma^2}{N-1}$$

- Inserendo l'espressione di cov appena trovata si ottiene:

$$\mathrm{var}(\overline{X}) = \frac{\sigma^2}{n} \cdot \frac{N-n}{N-1}$$

Media campionaria

$$\overline{X} = \frac{1}{n}(X_1 + X_2 + X_3 + \cdots + X_n)$$

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

$$VAR(\overline{X}) = \frac{\sigma^2}{n}$$

Varianza campionaria S^2

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

E(S^2) ?   VAR(S^2)?

Esercizio

Numero di esami sostenuti da 4 studenti del 1° anno di una certa facoltà

$$U = (2,\ 4,\ 4,\ 5)$$

Calcolare la distribuzione delle medie campionarie (n=2) nel caso di

-estrazione bernoulliana

-estrazione in blocco

Calcolare la distribuzione delle varianze campionarie (estrazione bernoulliana)

U = (2, 4, 4, 5)
Spazio dei campioni

μ = 3,75

σ^2 = 1,1875

	(2, 2)	**(2, 4)**	**(2, 4)**	**(2, 5)**
$\bar{x}$	2	3	3	3,5
s^2	0	1	1	2,25
	(4, 2)	**(4. 4)**	**(4, 4)**	**(4, 5)**
$\bar{x}$	3	4	4	4,5
s^2	1	0	0	0,25
	(4, 2)	**(4. 4)**	**(4, 4)**	**(4, 5)**
$\bar{x}$	3	4	4	4,5
s^2	1	0	0	0,25
	(5, 2)	**(5, 4)**	**(5, 4)**	**(5, 5)**
$\bar{x}$	3,5	4,5	4,5	5
s^2	2,25	0,25	0,25	0

U = (2, 4, 4, 5)
Spazio dei campioni

	(2, 2)	(2, 4)	(2, 4)	(2, 5)
$\bar{x}$	2	3	3	3,5
s^2	0	1	1	2,25
	(4, 2)	(4. 4)	(4, 4)	(4, 5)
$\bar{x}$	3	4	4	4,5
s^2	1	0	0	0,25
	(4, 2)	(4. 4)	(4, 4)	(4, 5)
$\bar{x}$	3	4	4	4,5
s^2	1	0	0	0,25
	(5, 2)	(5, 4)	(5, 4)	(5, 5)
$\bar{x}$	3,5	4,5	4,5	5
s^2	2,25	0,25	0,25	0

v. a. media campionaria

(estrazione Bernoulliana)

$\bar{x}_i$	n_i	p_i
2	1	0,0625
3	4	0,25
3,5	2	0,125
4	4	0,25
4,5	4	0,25
5	1	0,0625
	16	1

Universo U = (2, 4, 4, 5) $\mu = 3{,}75$ $\sigma^2 = 1{,}1875$

v. a. media campionaria

- **Estrazione bernoulliana**
- **E($\overline{X}$) = 3,75 = μ**
- **VAR($\overline{X}$) = $(2-3{,}75)^2$ 0,0625 +…+ $(5-3{,}75)^2$ 0,0625=0,594**

$$VAR(\overline{X}) = \frac{\sigma^2}{n} = \frac{1{,}1875}{2} = 0,594$$

$\overline{x}_i$	n_i	p_i
2	1	0,0625
3	4	0,25
3,5	2	0,125
4	4	0,25
4,5	4	0,25
5	1	0,0625
	16	1

Universo U = (2, 4, 4, 5) $\mu = 3{,}75$ $\sigma^2 = 1{,}1875$

Spazio dei campioni

	(2, 2)	(2, 4)	(2, 4)	(2, 5)
$\bar{x}$		3	3	3,5
s^2	0	1	1	2,25
	(4, 2)	(4, 4)	(4, 4)	(4, 5)
$\bar{x}$			4	4,5
s^2	1	0	0	0,25
	(4, 2)	(4, 4)	(4, 4)	(4, 5)
$\bar{x}$				4,5
s^2	1	0	0	0,25
	(5, 2)	(5, 4)	(5, 4)	(5, 5)
$\bar{x}$	5			
s^2	2,25	0,25	0,25	0

v. a. media campionaria

(estrazione in blocco)

$\bar{x}_i$	n_i	p_i
3	2	0,333
3,5	1	0,167
4	1	0,167
4,5	2	0,333
	6	1

Universo U = (2, 4, 4, 5) $\mu = 3{,}75$ $\sigma^2 = 1{,}1875$

v. a. media campionaria

- **Estrazione in blocco**

$\bar{x}_i$	n_i	p_i
3	2	0,333
3,5	1	0,167
4	1	0,167
4,5	2	0,333
	6	1

- **E($\bar{X}$) = 3,75 = μ**
- **VAR($\bar{X}$) = 0,396 =** $\frac{4-2}{4-1} \cdot \frac{1{,}1875}{2} = \frac{N-n}{N-1} \cdot \frac{\sigma^2}{n}$

Varianza campionaria S^2

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

- Calcolare teoricamente $E(S^2)$

$E(S^2)$

$$\sum_{i=1}^{n}(X_i-\mu)^2=\sum_{i=1}^{n}(X_i-\bar{X}_n+\bar{X}_n-\mu)^2=\sum_{i=1}^{n}(X_i-\bar{X}_n)^2+\sum_{i=1}^{n}(\bar{X}_n-\mu)^2$$

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

$$\mathrm{E}[S^2]=\sigma^2\frac{n-1}{n}$$

Universo U = (2, 4, 4, 5) $\mu = 3{,}75$ $\sigma^2 = 1{,}187$

Spazio dei campioni

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

	(2, 2)	**(2, 4)**	**(2, 4)**	**(2, 5)**
$\bar{x}$	2	3	3	3,5
s^2	0	1	1	2,25
	(4, 2)	**(4. 4)**	**(4, 4)**	**(4, 5)**
$\bar{x}$	3	4	4	4,5
s^2	1	0	0	0,25
	(4, 2)	**(4. 4)**	**(4, 4)**	**(4, 5)**
$\bar{x}$	3	4	4	4,5
s^2	1	0	0	0,25
	(5, 2)	**(5, 4)**	**(5, 4)**	**(5, 5)**
$\bar{x}$	3,5	4,5	4,5	5
s^2	2,25	0,25	0,25	0

v. a. varianza campionaria

(estrazione Bernoulliana)

	n_i	p_i
0	6	0,375
0,25	4	0,25
1	4	0,25
2,25	2	0,125
	16	1

Universo U = (2, 4, 4, 5) $\mu = 3,75$ $\sigma^2 = 1,187$

v.a. S^2 (varianza campionaria)

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

- Estrazione bernoulliana

S^2	n_i	p_i
0	6	0,375
0,25	4	0,25
1	4	0,25
2,25	2	0,125
	16	1

$$E(S^2) = 0{,}594 = 1{,}187\,\frac{1}{2} = \sigma^2\,\frac{n-1}{n}$$

$E(S^2) \neq 1,187 = \sigma^2$

Esercizio 2

- X = fenomeno dicotomico
  (A = acquirente, $\overline{A}$ = non acquirente)

U = ($A, A, \overline{A}, \overline{A}, \overline{A}$)

- Distribuzione di X? E(X)? VAR(X)?
- Spazio dei campioni (estrazione Bernoulliana) con n=2?
- Calcolo della distribuzione della frequenza relativa campionaria (P) e del numero di successi (S)? E(P)? VAR(P)? E(S)? VAR(S)?
- Distribuzione approssimata (asintotica) di P e S?

Distribuzione di X

- X = fenomeno dicotomico
  (A = acquirente, $\bar{A}$ = non acquirente)

$$U = (A, A, \bar{A}, \bar{A}, \bar{A})$$

x_i	p_i
$\bar{A}$ = 0	0,6
A = 1	0,4
	1

X = v. a. di bernoulli con π=0,4

$E(X) = \pi = 0{,}4$

$VAR(X) = \pi \cdot (1 - \pi) = 0{,}4 \cdot 0{,}6 = 0{,}24$

Gli elementi campionari $(X_1, \ldots, X_n)$ sono v.c. di Bernoulli con $E(X_i)=\pi=0{,}4$ e $VAR(X_i) = \pi \cdot (1 - \pi)=0{,}24$

$$U = (A, A, \overline{A}, \overline{A}, \overline{A})$$

- Spazio dei campioni (estrazione Bernoulliana) con n=2?
- Calcolo della distribuzione della frequenza relativa campionaria (P) e del numero di successi (S)? E(P)? VAR(P)? E(S)? VAR(S)?

$$U = (A, A, \bar{A}, \bar{A}, \bar{A})$$

π =2/5=0,4

Spazio dei campioni (n=2)

$$\begin{matrix} (A,A) & (A,A) & (A,\bar{A}) & (A,\bar{A}) & (A,\bar{A}) \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ (\bar{A},A) & (\bar{A},A) & (\bar{A},\bar{A}) & (\bar{A},\bar{A}) & (\bar{A},\bar{A}) \end{matrix}$$

Numero dei possibili campioni: 5^2=25

Variabile aleatoria *P*

$$P = \frac{1}{n}(X_1 + X_2 + ... + X_n)$$

s	s/n	p(s)
0	0	9/25= $0,6^2$=0,36
1	0,5	12/25 = 2·0,4·0,6=0,48
2	1	4/25 = $0,4^2$ = 0,16
		1

P(s) = $\binom{n}{s}\pi^s \cdot (1-\pi)^{n-s}$

Caratteristiche della v.a. P
(e confronto con v.a. media campionaria)

$$\boldsymbol{P = \frac{1}{n}(X_1 + X_2 + \ldots + X_n)}$$

$$\bar{X} = \frac{1}{n}(X_1 + X_2 + \ldots + X_n)$$

$$E(P) = \pi$$

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

$$VAR(P) = \frac{\pi \cdot (1-\pi)}{n}$$

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

- **Forma di distribuzione**
  - **Esatta ⇒ binomiale: P=S/n ~ B(n, π)/n**
  - **Se n>100 si può applicare il Teorema centrale del limite:**

$$P \sim N\left(\pi, \frac{\pi \cdot (1-\pi)}{n}\right)$$

Distribuzione della v.a. frequenza relativa di successi

$$P \sim N\left(\pi, \frac{\pi \cdot (1-\pi)}{n}\right)$$

Distribuzione della v.a. numero di successi

Anche

X = n. di successi = *n* × *P*

per n > 100

X~ N(n·π, n·π(1-π))

Esempio

In una città i contribuenti che hanno presentato il modello 730 sono stati 12.000 ed il reddito medio è stato di 15.000 euro, con s.q.m. pari a 8.000 euro.

Estraendo da tale universo un campione casuale di 200 contribuenti si determini la probabilità che la media campionaria sia compresa tra 14.000 e 17.000 euro:

- **nell'ipotesi di estrazione con reimmissione (bernoulliana)**
- **nel caso di estrazione senza reimmissione**

Si illustrino inoltre le assunzioni che rendono possibile il calcolo.

Richiamo distribuzione della v.a. media campionaria

$$\frac{\bar{X} - E(\bar{X})}{\sqrt{\mathrm{var}(\bar{X})}} \xrightarrow{Legge} \mathrm{N}(0,1)$$

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

$$\bar{X} - \mu \sim \mathrm{N}(0, \sigma^2 / n)$$

$$\bar{X} \sim \mathrm{N}(\mu, \sigma^2 / n)$$

Distribuzione v.a. media campionaria (n elevato)

- Estrazione con reimmissione

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

- Estrazione senza reimmisione

$$\bar{X} \sim \mathrm{N}\left(\mu, \frac{\sigma^2}{n} \cdot \frac{N-n}{N-1}\right)$$

SOLUZIONE

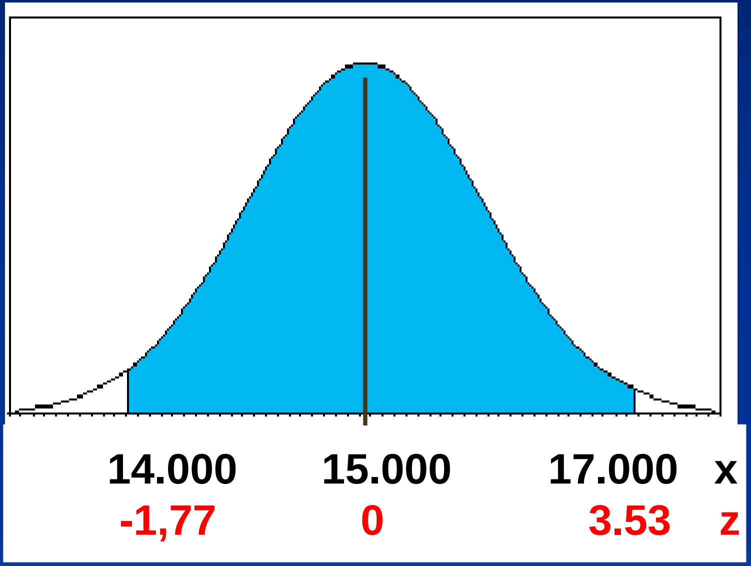
$\mu = 15.000 \quad \sigma = 8.000 \quad n = 200 \quad N = 12.000$

a) $Pr(14000 < \overline{X} < 17000)$

$$\overline{X} \sim \mathrm{N}(\mu, \sigma^2 / n)$$

$$\sigma(\overline{X}) = \frac{\sigma}{\sqrt{n}} = \frac{8.000}{\sqrt{200}} = 565,68$$

$$\overline{X} \approx N(15.000; 565,68^2)$$



$$z_1 = \frac{14.000 - 15.000}{565,68} = -1,77; \qquad z_2 = \frac{17.000 - 15.000}{565,68} = 3,53$$

$F(z_2) - F(z_1) = 0,99979 - 0,03836 = 0,96143$

b) Estrazione in blocco

$$\overline{X} \sim \mathrm{N}\left(\mu, \frac{\sigma^2}{n} \cdot \frac{N-n}{N-1}\right)$$

$$\sigma(\overline{X}) = \frac{8.000}{\sqrt{200}} \cdot \sqrt{\frac{12.000-200}{12.000-1}} = 560,9749$$

$$\overline{X} \approx N(15.000; 560,9749^2)$$

$$\Pr(14000 < \overline{X} < 17000)$$

$$z_1 = \frac{14.000-15.000}{560,9749} = -1,78; \qquad z_2 = \frac{17.000-15.000}{560,9749} = 3,57$$

- $F(z_2) - F(z_1) = 0,99982 - 0,03754 = 0,9623$

Confronto prob intervallo nel caso di estrazione in blocco e con reimmissione

b) Estrazione Bernoullina

$$\Pr(14000 < \bar{X} < 17000) = 0{,}96143$$

b) Estrazione in blocco

$$\Pr(14000 < \bar{X} < 17000) = 0{,}9623$$

Esempio

- **Un test per l'ammissione ad un corso di laurea a numero chiuso consiste in 100 domande, per ciascuna delle quali sono indicate 4 possibili risposte, di cui una sola esatta. Un candidato assolutamente impreparato, risponde semplicemente barrando a caso una risposta per ogni domanda.**
- **Qual è la probabilità che egli fornisca almeno 40 risposte esatte? (Si indichi dapprima l'espressione esatta per il calcolo di tale probabilità e la si determini poi avvalendosi di un teorema appropriato)**

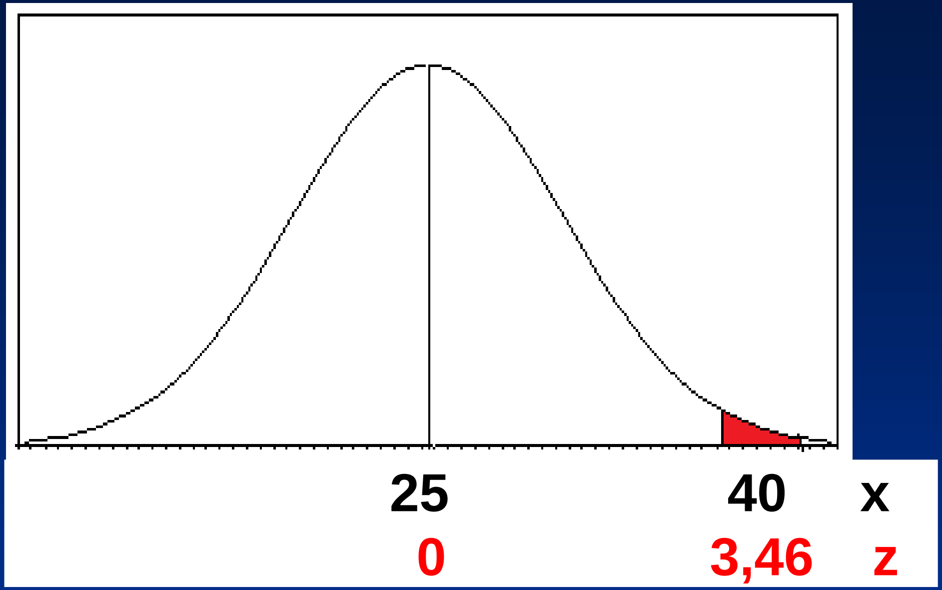
Soluzione

n =100 π= ¼ = 0,25 S~B(100; 0,25)

- **Pr (S≥40) =** $$\sum_{s=40}^{100} \binom{100}{s} \cdot 0{,}25^{s} \cdot (1-0{,}25)^{100-s} = 0{,}000687$$
- **Teorema Centrale del limite:**
  Z(S) ~ N(0, 1) n→ ∞ E(S)=? VAR(S)=?
- **E(S) = n π = 25**
- **VAR(S) = n π (1-π) = 18,75**

$$S \approx N(25;18{,}75)$$

$$S \approx N(25;18,75)$$



$$z = \frac{40-25}{\sqrt{18,75}} = 3,46$$

Pr(S≥40) = Pr(z≥3,46) = 1 – F(3,46) =

1 – 0,99973 = 0,00027

La scelta degli elementi campionari

pp. 47-51

Le origini delle indagini campionarie

- 1936: Elezioni presidenziali USA
- Candidati: F.D. Roosvelt e A. Landon
- Indagine Literary Digest:
  - 10 milioni di fac-simile di schede elettorali inviate a nominativi estratti dagli elenchi telefonici e dai registri automobilistici
  - Risultato previsto: Roosvelt 41% e Landon 59%
- Indagine Gallup:
  - alcune migliaia di interviste ad elettori estratti casualmente dall'intera popolazione
  - Risultato previsto: Roosvelt 60% e Landon 40%

Risultato delle elezioni: Roosvelt 61%

- Gli errori del Literary Digest:
- ERRORE DI COPERTURA
  - le liste usate non erano complete
  - gli elenchi usati non erano rappresentativi dell'intera popolazione ma solo dei ceti più abbienti che tendevano a votare repubblicano
- AUTOSELEZIONE del CAMPIONE
  - Le caratteristiche socio-demograche dei cittadini che risposero al sondaggio erano presumibilmente diverse da quelle di chi non rispose (istruzione, reddito, etc.)

Problemi operativi

DALL'UNIVERSO SI ESTRE UN UNICO CAMPIONE:

- **COME SI SCELGONO LE UNITA' CAMPIONARIE?**

  ⇒ Un campione dovrebbe essere *rappresentativo* dell'universo da cui è tratto

TIPI DI CAMPIONAMENTO

Problemi operati

- QUANTE DEVONO ESSERE LE UNITA' CAMPIONARIE?

$\Rightarrow$ determinazione di n:

  - Limiti di costo
  - Precisione delle stime
  - Indagini di natura sociale e nell'ambito delle scienze sperimentali

TIPI DI CAMPIONAMENTO

- **Campione probabilistico $\Rightarrow$ ogni elemento dell'universo ha una probabilità prefissata di comparire nel campione**

- **Campione non probabilistico $\Rightarrow$**

(non è possibile estendere all'universo in termini di probabilità le informazioni ottenute dal campione)

Campionamento non probabilistico

- A scelta ragionata (sceglie le persone da intervistare che meglio rispondono alle finalità della propria indagine)
- A valanga (Consiste nell'individuare un primo gruppo di persone da intervistare e che possiedono le caratteristiche cercate. Questi soggetti sono a loro volta "invitati" (utilizzati) per individuare altri soggetti con le stesse caratteristiche

Campionamento non probabilistico

- Di convenienza (“cheap and dirty”)

Campionamento casuale semplice (con o senza reimmissione)

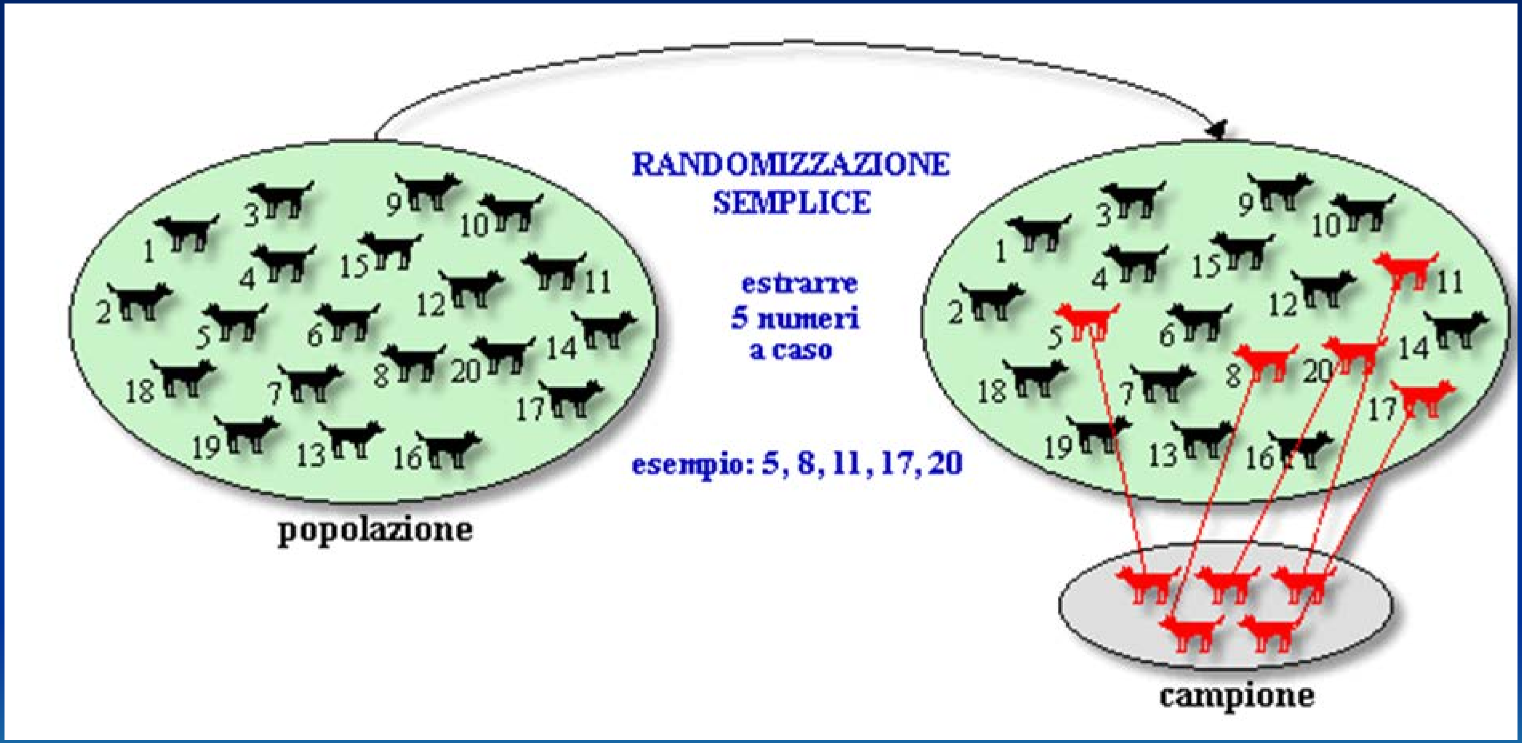
- **Ogni elemento dell'universo ha la medesima probabilità di comparire nel campione**

Scelta in concreto:

- **lista delle N unità dell'universo**
- **numeri aleatori (tavole o generatori casuali)**

Es. di camp. casuale semplice

- Es. popolazione di cani.



- Il principale svantaggio è quello di richiedere la preventiva numerazione di tutte le unità; successivamente è necessario individuare nella popolazione quelli corrispondenti ai numeri estratti

Campionamento sistematico

- Il campione è costruito selezionando un elemento dalla lista di campionamento ogni *k elementi.*
- Il campionamento sistematico si può applicare sia ad una popolazione già numerata progressivamente (elenco telefonico, produzione di oggetti consecutivi)
- Come si sceglie k? In base alla porzione di popolazione che mi interesserà intervistare.

Es. di camp. sistematico

- Es. popolazione di cani. Si seleziona un animale ogni 4

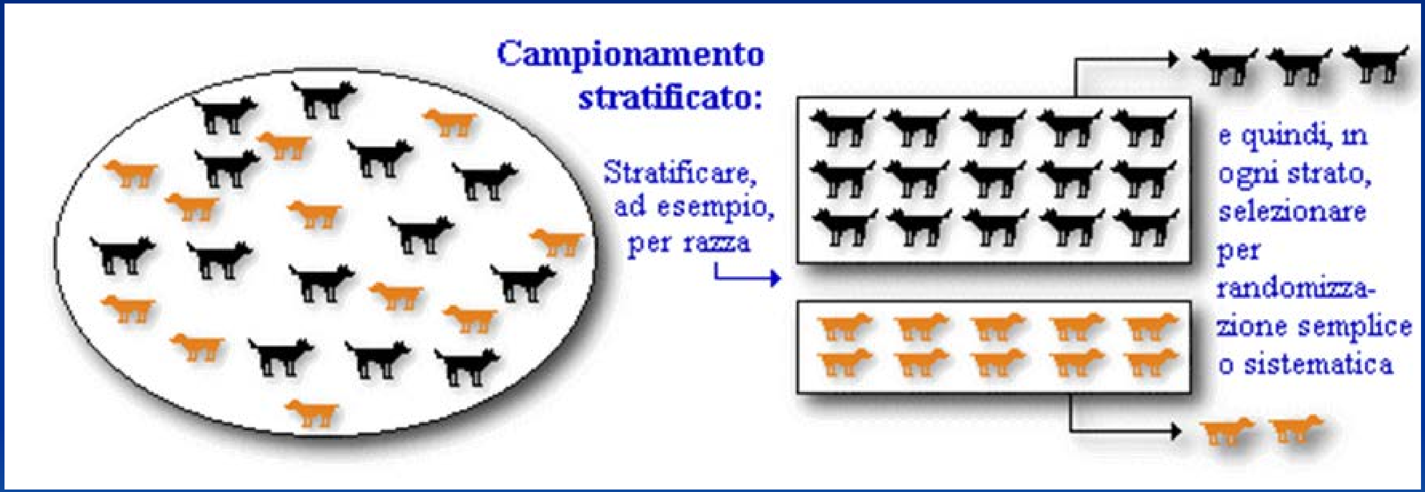
- Il campionamento sistematico deve essere usato con cautela quando la successione degli elementi nell'universo presenta delle regolarità cicliche

Campionamento stratificato

- Tiene conto di informazioni supplementari (note a priori) sulla popolazione → suddivisa in gruppi omogenei (strati)
- Si estrae un campione casuale da ogni strato, secondo una proporzione stabilita (spesso strat. Proporzionale)
- A parità di dimensione campionaria→ maggiore precisione (tutti gli strati sono rappresentati nel campione)

Es. di camp. stratificato

- Es. popolazione di cani. Si stratifica per razza

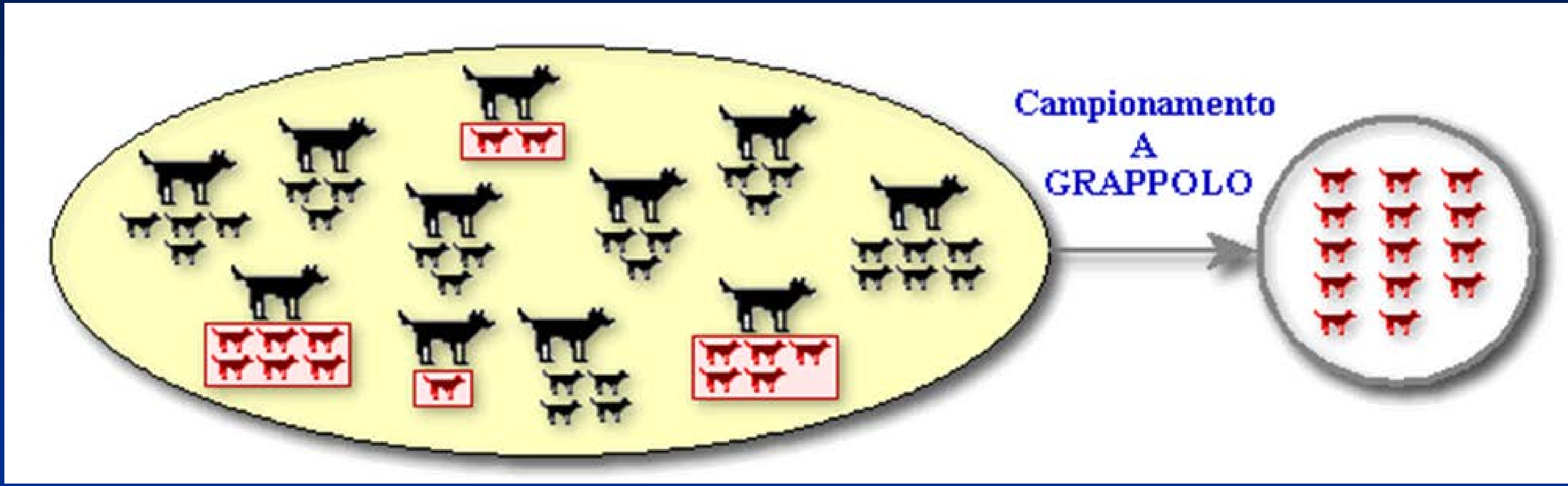


- Un campione ottenuto per stratificazione ha il vantaggio di **rappresentare meglio** la popolazione da cui è stato estratto; tuttavia, la ridotta numerosità dei vari strati può rendere poco attendibili le stime riferite ai singoli strati.

Campionamento a grappolo

- La popolazione viene suddivisa in gruppi che diventano le unità su cui si effettua il campionamento
- ESEMPIO. Si deve stimare la presenza di una malattia che colpisce i cuccioli di cane poco dopo la nascita. L'unità di indagine è rappresentata dal «cucciolo». Si procede ad effettuare un campionamento a grappolo, selezionando, mediante randomizzazione semplice o sistematica, un certo numero di *nidiate*.

Campionamento a grappolo



- Rispetto alla randomizzazione semplice, sistematica o stratificata, il campionamento a grappolo offre il vantaggio di facilitare notevolmente il reclutamento dei soggetti; di conseguenza si abbassano costi e tempi dell'indagine

Campionamento a stadi

- Si individuano le unità di primo stadio (es. comuni)
- Si estraggono elementi da ogni stadio
- Es. comuni (unità di primo stadio) → famiglie
- Es. scuole (unità di primo stadio) → studenti