

Il campione

- Prima definizione
 - È una parte del tutto (la popolazione)
 - Non è detto che sia così:
 - Operazione di estrazione
 - Con ripetizione (può non essere un sottoinsieme)
 - Senza ripetizione
 - In blocco
 - Un'unità per volta

Il campione

- Altre definizioni

- Sia una popolazione di N unità $P = \{1, 2, \dots, N\}$
 - Si estraggono da questa n unità in modo sequenziale
 - con o senza ripetizione
 - Sia i_1 l'etichetta della prima unità estratta
 - Sia i_2 l'etichetta della seconda unità estratta
 - ...
 - Sia i_n l'etichetta dell'ennesima unità estratta
- Le n estrazioni generano una sequenza di indici

$$S = \{i_1, i_2, \dots, i_n\}$$

Il campione

- Si chiama ***campione di dimensione n*** della popolazione P un qualsiasi sottoinsieme o sequenza

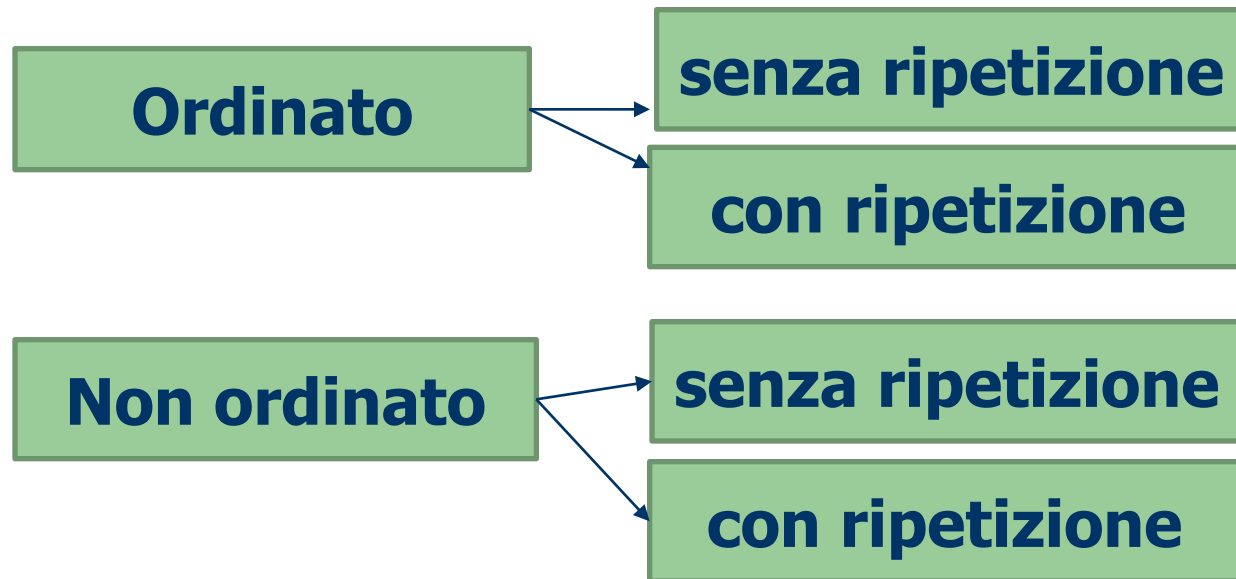
$$S = \{i_1, i_2, \dots, i_n\}$$

di P contenente n unità

i_j	$j=1, \dots, n$	etichetta dell'unità i -esima scelta alla j -esima estrazione
-------	-----------------	--

Il campione

- Il ***campione di dimensione n*** della popolazione P può essere



Campione ordinato

- Esempio .. se si tiene conto dell'ordine
 $N=10$ unità $P = \{1,2,3,4,5,6,7,8,9,10\}$
 - Si supponga di estrarre “con ripetizione” un campione ordinato di $n=3$ unità
 - $i_1=9$ $i_2=3$ $i_3=3$
 - Il campione è identificato dalla sequenza (9,3,3) ed è diverso dalla sequenza (3,9,3)
 - $i_1=3$ $i_2=9$ $i_3=3$

Campione ordinato

- Esempio .. se si tiene conto dell'ordine
N=10 unità
 - Si supponga di estrarre “senza ripetizione” un campione ordinato di n=3 unità
 $i_1=9 \ i_2=3 \ i_3=2$ e $i_2 \neq 9 \ i_3 \neq 9$ e $i_3 \neq 3$
 - Il campione è identificato dalla sequenza (9,3,2) ed è diverso dalla sequenza (2,9,3)
 $i_1=2 \ i_2=9 \ i_3=3$ e $i_2 \neq 2 \ i_3 \neq 2$ e $i_3 \neq 9$

Campione non ordinato

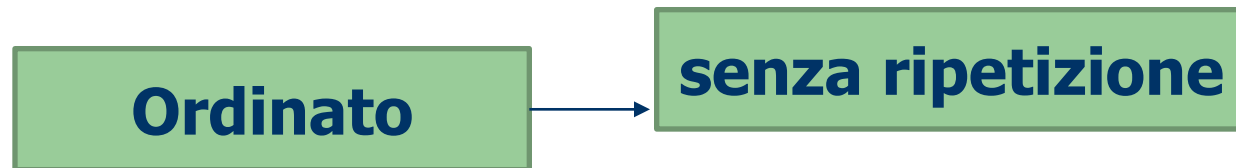
- Esempio .. se non si tiene conto dell'ordine
 $N=10$ unità $P = \{1,2,3,4,5,6,7,8,9,10\}$
 - Si supponga di estrarre “con ripetizione” un campione non ordinato di $n=3$ unità
 - $i_1=9$ $i_2=3$ $i_3=3$
 - Il campione è identificato dal sottoinsieme (9,3,3) ed è uguale al sottoinsieme (3,9,3)
 - $i_1=3$ $i_2=9$ $i_3=3$

Campione non ordinato

- Esempio .. se non si tiene conto dell'ordine
N=10 unità
 - Si supponga di estrarre “senza ripetizione” un campione non ordinato di $n=3$ unità
 $i_1=9$ $i_2=3$ $i_3=2$
 - Il campione è identificato dal sottoinsieme (9,3,2) ed è uguale al sottoinsieme (2,9,3)
 $i_1=2$ $i_2=9$ $i_3=3$

Il campione

Numero di ***campioni distinti*** a seconda della tipologia

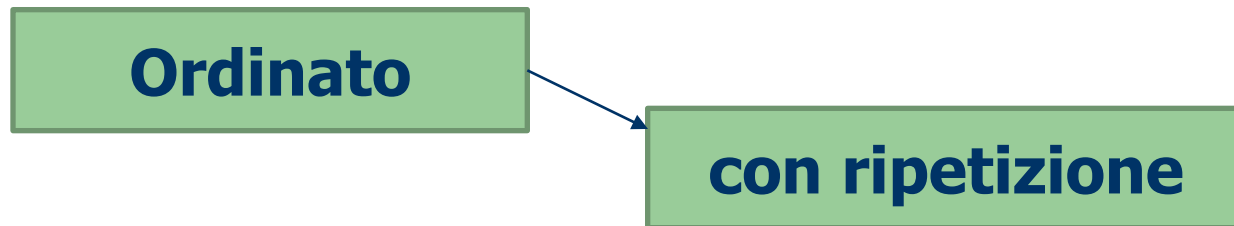


$$D_{N,n} = N(N-1)(N-2)\dots(N-n+1)$$

Disposizioni senza ripetizione di classe n

Il campione

- Numero di ***campioni distinti*** a seconda della tipologia



Disposizioni con ripetizione di classe n

$$N^n$$

Il campione

- Numero di ***campioni distinti*** a seconda della tipologia

$$\binom{N}{n}$$

Combinazioni di n elementi da N

Non ordinato

senza ripetizione

Il campione

- Numero di ***campioni distinti*** a seconda della tipologia

$$\binom{N+n-1}{n}$$

Combinazioni di n elementi da $N+n-1$

Non ordinato

con ripetizione

Il campione

- Qual è il campione più “naturale”?
 - “senza ripetizione” o “con ripetizione”?
 - Ordinato o non ordinato?

Non ordinato e senza ripetizione

Ricapitolando ...

Campioni ordinati (sequenze)

con ripetizione: N^n

senza ripetizione: $D_{N,n} = N(N-1)(N-2)\dots(N-n+1)$

Se ho una popolazione di $N=10$ elementi e devo estrarre un campione di $n=3$ elementi lo spazio campionario sarà costituito da:

$N^n = 10^3 = 1000$ campioni diversi estratti con ripetizione

$D_{N,n} = N(N-1)(N-2)\dots(N-n+1) = 10 \times 9 \times 8 \times 7 \times 6 \times 5 \times 4 = 30.240$

Ricapitolando ...

Campioni non ordinati (sottoinsiemi)

con ripetizione: $C'_{N,n} = \binom{N+n-1}{n}$

senza ripetizione: $C_{N,n} = \binom{N}{n}$

Se ho una popolazione di $N=10$ elementi e devo estrarre un campione di $n=3$ elementi lo spazio campionario sarà costituito da:

$$C_{N,n} = \binom{N}{n} = \binom{10}{3} = \frac{10 \times 9 \times 8}{3 \times 2} = 120 \quad \text{campioni non ordinati senza ripetizione}$$

Dimensione campionaria

In una sequenza il numero di componenti di s è chiamata dimensione campionaria

$$n(S)$$

Se le estrazioni sono con ripetizione e sequenzialmente effettuate

$n(S)$ può essere maggiore di N

L'effettiva dimensione campionaria

$v(S)$ è rappresentata

dal numero dei componenti distinti in una sequenza
quando in $n(S)$ gli elementi sono tutti distinti

$$n(S) = v(S)$$

Dimensione campionaria

Estrazioni senza ripetizione

il campione è costituito da unità elementari differenti:

$$S \subset P \text{ e } n \leq N$$

Estrazioni con ripetizione

il campione può essere costituito da unità elementari già presenti:

$$S \not\subset P \text{ e si può verificare che } n > N$$

Lo spazio campionario

- Definizione:

- Lo spazio campionario è l'insieme di tutte le sequenze o di tutti i sottoinsiemi e si indica con S (o S^* o Ω o Ω^*)
- Lo spazio campionario è l'insieme di tutti i possibili campioni di dimensione n che si possono estrarre da una popolazione di dimensioni N , in base ad una tecnica predefinita
- Numero di possibili campioni?
 - Dipende dall'ordine e dalla tecnica di estrazione

Lo spazio campionario

- Si può considerare anche il numero di *campioni non ordinati* di ampiezza variabile
- In questo caso il piano di campionamento viene indicato con Ω^*

Lo spazio campionario

- Il numero di *campioni non ordinati* può essere ottenuto come
 - Riduzione dall'insieme di campioni ordinati
 - S^* è lo spazio campionario dei campioni ordinati
 - S è lo spazio campionario dei campioni non ordinati

$$\Omega_{n.ord} < \Omega_{ord}^* \rightarrow S < S^*$$

Lo spazio campionario

– Esempio

- P (1,2,3) n=1,2,3 senza ripetizione

$$s_1=1$$

$$s_4=1,2$$

$$s_{10}=1,2,3$$

$$s_2=2$$

$$s_5=2,1$$

$$s_{11}=2,3,1$$

$$s_3=3$$

$$s_6=1,3$$

$$s_{12}=3,1,2$$

$$s_7=3,1$$

$$s_{13}=1,3,2$$

$$s_8=2,3$$

$$s_{14}=2,1,3$$

$$s_9=3,2$$

$$s_{15}=3,2,1$$

Lo spazio campionario che tiene conto dell'ordine sarà

$$\Omega_{ord} = \{s_1, s_2, \dots, s_{15}\}$$

Lo spazio campionario senza ordine, invece, sarà

$$\Omega_{n.ord}^* = \{s_1, s_2, s_3, s_4, s_6, s_8, s_{10}\}$$

Il Piano di campionamento

Associamo ad ogni campione s una misura di probabilità

$$p(s) \geq 0; \sum_s p(s) = 1$$

■ Definizione:

- Il piano di campionamento è una funzione $p(s)$ su S che soddisfa le due relazioni precedenti

Il Piano di campionamento

- Per gli spazi campionari descritti si hanno i seguenti piani di campionamento:
 - Piano di campionamento per campioni ordinati senza ripetizione

$$p(s) = \frac{1}{D_{N,n}} = \frac{1}{N(N-1)(N-2)\dots(N-n+1)}$$

- Piano di campionamento per campioni ordinati con ripetizione

$$p(s) = \frac{1}{D_{N,n}} = \frac{1}{N^n}$$

Il Piano di campionamento

- Per gli spazi campionari descritti si hanno i seguenti piani di campionamento:
 - Piano di campionamento per campioni non ordinati senza ripetizione

$$p(s) = \frac{1}{C_{N,n}} = \frac{1}{\binom{N}{n}}$$

- Piano di campionamento per campioni non ordinati con ripetizione

$$p(s) = \frac{1}{C_{N,n}} = \frac{1}{\binom{N+n-1}{n}}$$

Il Piano di campionamento

.. Processo di selezione e stima dei parametri incogniti della popolazione

- Varia al variare dell'obiettivo

- Deve essere misurabile

- Errori standard
- Funzioni di verosimiglianza
- Distribuzioni di campionamento

- Analisi di fattibilità economica

- raggiungimento degli obiettivi al minimo costo (denaro/tempo)

- Cura nel tradurre un modello teorico di selezione in un insieme di istruzioni utili per gli esecutori

- Chiare, Semplici, Pratiche, Complete

Probabilità di inclusione

Probabilità di inclusione

La probabilità che una unità (o un gruppo di unità) appartenga al campione estratto

Probabilità di inclusione del primo ordine

- Si consideri la generica unità i della popolazione P
- Sia A_i l'insieme dei campioni dello spazio campionario S (o S^*) che contengono l'unità i

La probabilità di inclusione del primo ordine dell'unità i π_i

è data dalla somma delle probabilità dei campioni appartenenti ad A_i

$$\pi_i = \sum_{A_i} p(s)$$

Probabilità di inclusione

Esempio:

Sia P una popolazione $P(1,2,3,4)$

Costruiamo i campioni di $n=2$

non ordinati senza ripetizione

Probabilità
associate ai
campioni

$$s \begin{bmatrix} (1,2) & (1,3) & (1,4) & (2,3) & (2,4) & (3,4) \\ p(s) & 0,15 & 0,10 & 0,20 & 0,15 & 0,20 & 0,20 \end{bmatrix}$$

Probabilità di inclusione

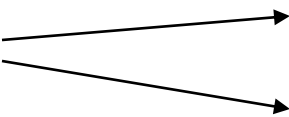
$$\pi_1 = 0,15 + 0,10 + 0,20 = 0,45$$

$$\pi_2 = 0,15 + 0,15 + 0,20 = 0,50$$

$$\pi_3 = 0,10 + 0,15 + 0,20 = 0,45$$

$$\pi_4 = 0,20 + 0,20 + 0,20 = 0,60$$

$$\pi_i = \sum_s \delta_i p(s) = E(\delta_i)$$

δ_i  1 se l'unità i è inclusa nel campione
0 altrimenti

Probabilità di inclusione

Probabilità di inclusione del secondo ordine

- Si concentri l'attenzione su una particolare coppia di unità (i e j) della popolazione P .
- Sia A_{ij} l'insieme dei campioni dello spazio campionario S (o S^*) che contengono le unità i e j

La probabilità di inclusione del secondo ordine delle unità i e j

$$\pi_{ij}$$

è data dalla somma delle probabilità dei campioni appartenenti ad A_{ij}

$$\pi_{ij} = \sum_{A_{ij}} p(s)$$

Probabilità di inclusione


Esempio:

Sia P una popolazione $P(1,2,3,4,5)$

Costruiamo i campioni di $n=4$

non ordinati senza ripetizione

Probabilità
associate ai
campioni



s	$(1,2,3,4)$	$(1,2,3,5)$	$(1,2,4,5)$	$(2,3,4,5)$	$(1,3,4,5)$
$p(s)$	0,15	0,25	0,10	0,30	0,20

Probabilità di inclusione

$$\pi_{12} = 0,15 + 0,25 + 0,10 = 0,50$$

$$\pi_{13} = 0,15 + 0,25 + 0,20 = 0,60$$


$$\pi_{14} = 0,15 + 0,10 + 0,20 = 0,45$$


$$\pi_{23} = 0,15 + 0,25 + 0,30 = 0,70$$

$$\pi_{24} = 0,15 + 0,10 + 0,30 = 0,55$$

$$\pi_{34} = 0,15 + 0,30 + 0,00 = 0,65$$

$$\pi_{ij} = \sum_s \delta_i \delta_j p(s) = E(\delta_i \delta_j)$$

δ_i  1 se le unità i e j sono incluse nel campione

δ_j  0 se i o se j o se i e j non sono incluse nel campione

Probabilità di inclusione

Se il campionamento è **con ripetizione** ci si chiede qual è il numero di volte (in media) che una singola unità si presenta nel campione

Sia γ_i il numero di volte che l'unità i appare nel campione S .

In un campione di dimensione n tale variabile può assumere i valori $\{0,1,2,3,\dots,n\}$

La quantità $\phi_i = \sum_s \gamma_i p(s) = E(\gamma_i)$

È la frequenza attesa di inclusione

$$\pi_i = \phi_i$$

Se il campionamento è senza ripetizione $\delta_i = \gamma_i$

Probabilità di inclusione

Un piano campionario è autoponderante se tutte le unità della popolazione hanno la

stessa probabilità di inclusione del primo ordine

o la

stessa frequenza attesa di inclusione

Si definisce probabilità di inclusione di ordine k la probabilità che k unità prestabilite appartengano contemporaneamente al campione estratto

Strategia campionaria

.. Si intende il binomio

Piano di campionamento

+

Stimatore

Dato un piano di campionamento vi è una pluralità di stimatori possibili

Come scegliere ?

- a) *in base alle proprietà degli stimatori*
- b) *per l'effetto del disegno (Kish, 1965)*

Stimatori e effetto del disegno

La scelta degli stimatori viene effettuata prima di definire il piano dell'indagine, ma può essere rinviata alla fase di analisi dei dati.

- Quando lo stimatore scelto è corretto, la strategia campionaria si definisce corretta
- quando lo stimatore è efficiente la strategia campionaria si dirà efficiente

Proprietà degli stimatori: distorsione

$$\hat{\theta} = f(X_1, X_2, \dots, X_n)$$

è uno stimatore non distorto di θ se ha la distribuzione campionaria con media uguale al parametro da stimare. La distorsione non è altro che la differenza fra il valor medio dello stimatore ed il valor del parametro da stimare

$$\text{Distorsione} = B(\hat{\theta}) = E(\hat{\theta}) - \theta$$

La distorsione può essere positiva o negativa, ovvero la differenza maggiore o minore di 0.

$$B(\hat{\theta}) = 0 \quad \text{se} \quad E(\hat{\theta}) = \theta$$

Stimatori e effetto del disegno

Proprietà degli stimatori: Efficienza

Uno stimatore $\hat{\theta}$ è efficiente se:

$$E(\hat{\theta}) = \theta \quad E(\hat{\theta}_1) = \theta$$
$$Var(\hat{\theta}) < Var(\hat{\theta}_1)$$

dove

$\hat{\theta}_1$ è qualsiasi altro stimatore non distorto.

nella classe degli stimatori non distorti lo stimatore efficiente è quello che ha la varianza minima

Stimatori e effetto del disegno

Proprietà degli estimatori: Efficienza

è possibile fare riferimento ad una grandezza che tiene conto sia della distorsione che della (efficienza) varianza

$$MSE(\hat{\theta}) = E(\hat{\theta} - \theta)^2 = \text{var}(\hat{\theta}) + B^2(\hat{\theta})$$

Errore quadratico medio – misura la precisione delle stime

permette di definire l'efficienza anche per gli estimatori distorti, infatti se due o più estimatori sono distorti, lo stimatore efficiente è quello che ha il più piccolo M.S.E.

$$EF\left(\frac{\hat{\theta}_1}{\hat{\theta}_2}\right) = \frac{MSE(\hat{\theta}_1)}{MSE(\hat{\theta}_2)} \quad \text{se} \quad EF < 1 \quad \hat{\theta}_1 \longrightarrow \text{più efficiente di } \hat{\theta}_2$$

1. con n grande si preferiscono estimatori non distorti
2. fra due estimatori non distorti si sceglie quello più efficiente (con MSE più piccolo)

Stimatori e effetto del disegno

Effetto del disegno

Dato uno stimatore non distorto $\hat{\theta}$ di θ

ed un piano di campionamento, si chiama effetto del disegno il rapporto

$$Deff = \frac{Var(\hat{\theta})}{Var_0(\hat{\theta})}$$

Dove – a parità di n – il denominatore coincide con la varianza nel caso di campionamento casuale semplice (proprio perché il più semplice).

Al numeratore la varianza dello stimatore nel piano di campionamento scelto.

Il Deff è maggiore di 1 se la precisione dello stimatore nel piano di campionamento considerato è minore della stessa nel campionamento casuale semplice ovvero:

$$v(\hat{\theta}) > v_0(\hat{\theta})$$

Il Deff è minore di 1 se la precisione dello stimatore nel piano di campionamento considerato è maggiore della stessa nel campionamento casuale semplice:

$$v(\hat{\theta}) < v_0(\hat{\theta})$$

Notazioni sulla popolazione

- $P = \text{popolazione finita}$
 - Una popolazione finita P è una raccolta di N unità, $N < \infty$
 - $1, 2, 3, \dots, N$ “*etichette*”

Y Caratteristica oggetto di studio (Y_1, Y_2, \dots, Y_N)

unità
valori di Y $\begin{bmatrix} 1 & 2 & \cdot & \cdot & N \\ Y_1 & Y_2 & \cdot & \cdot & Y_N \end{bmatrix} \rightarrow \text{distribuzione semplice}$

Notazioni sulla popolazione

- Caso più frequente

valori di Y $\begin{bmatrix} Y_1 & Y_2 & \cdot & \cdot & Y_k \end{bmatrix}$
frequenze $\begin{bmatrix} N_1 & N_2 & \cdot & \cdot & N_k \end{bmatrix} \rightarrow$ distribuzione di frequenza

$$\sum_{h=1}^K N_h = N = N_1 + N_2 + \dots + N_k$$

Notazioni sulla popolazione

- Se osserviamo due variabili X e Y
 - *Distribuzione doppia*

$$\begin{array}{l} \text{unità} \\ \text{valori di X} \\ \text{valori di Y} \end{array} \begin{bmatrix} 1 & 2 & . & . & N \\ X_1 & X_2 & & & X_N \\ Y_1 & Y_2 & & & Y_N \end{bmatrix} \rightarrow \text{distribuzione doppia}$$

Notazioni sulla popolazione

- Caso più frequente
 - *Tabella a doppia entrata*

	Y_1	Y_2	\cdot	Y_k	
X_1	N_{11}	N_{12}	\cdot	N_{1k}	$N_{1\cdot}$
X_2	N_{21}	N_{22}	\cdot	N_{2k}	$N_{2\cdot}$
\cdot	\cdot	\cdot	\cdot		\cdot
X_h	N_{h1}	N_{h2}	\cdot	N_{hk}	$N_{h\cdot}$
	$N_{\cdot 1}$	$N_{\cdot 2}$	\cdot	$N_{\cdot k}$	N

Notazioni su costanti caratteristiche di popolazione

- Media

$$\bar{Y} = \left| \begin{array}{l} \frac{1}{N} \sum_{i=1}^N Y_i \text{ se } Y \text{ è quantitativa} \\ \frac{1}{N} \sum_{i=1}^N Y_i = \frac{N_A}{N} \text{ se } Y \text{ è qualitativa dicotoma} \end{array} \right.$$

$Y_i = 1$ se $i \in$ alla caratteristica A

$$\sum Y_i = N_A$$

Notazioni su costanti caratteristiche di popolazione

- Totale $\hat{Y} = N\bar{Y} = \begin{cases} \sum_{i=1}^N Y_i & \text{se } Y \text{ è quantitativa} \\ \sum_{i=1}^N Y_i = N_A & \text{se } Y \text{ è qualitativa dicotoma} \end{cases}$

N_A = Numero di unità che presentano la modalità A

Notazioni su costanti caratteristiche di popolazione

- Varianza $\sigma^2 = \frac{1}{N} \sum (Y_i - \bar{Y})^2$
- Deviazione standard $\sigma = \left[\frac{1}{N} \sum (Y_i - \bar{Y})^2 \right]^{\frac{1}{2}}$
- Coefficiente di variazione $CV = \frac{\sigma}{\bar{Y}}$
- Momento centrale $\bar{\mu}_r = \frac{1}{N} \sum (Y_i - \bar{Y})^r$

r=2 → varianza

Notazioni su costanti caratteristiche di popolazione

- Momento non centrale $\mu_r = \frac{1}{N} \sum (Y_i - k)^r$
- Con $K=0$ $\mu_r = \frac{1}{N} \sum (Y_i)^r$

$r=1$ \longrightarrow media aritmetica

Notazioni su costanti caratteristiche di popolazione

- Asimmetria

$$\gamma_1 = \frac{\overline{\mu_3}}{\sigma^3} < 0 \rightarrow \text{asimmetria negativa}$$
$$\gamma_1 = \frac{\overline{\mu_3}}{\sigma^3} > 0 \rightarrow \text{asimmetria positiva}$$

- Curtosi

$$\gamma_2 = \frac{\overline{\mu_4}}{\sigma^4} < 0 \rightarrow \text{iponormalità}$$
$$\gamma_2 = \frac{\overline{\mu_4}}{\sigma^4} > 0 \rightarrow \text{ipernormalità}$$

Notazioni su costanti caratteristiche di popolazione

- Distribuzioni doppie

- Rapporto fra medie $\frac{\bar{X}}{\bar{Y}}$

- Rapporto fra totali $\frac{X}{Y} = \frac{N\bar{X}}{N\bar{Y}}$

coincide
con il primo

- Covarianza $\sigma_{xy} = \frac{1}{N} \sum (X_i - \bar{X})(Y_i - \bar{Y})$

Se frequenza = 1

Notazioni su costanti caratteristiche di popolazione

- Distribuzioni doppie
 - Coefficiente di correlazione $\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$
 - Coefficiente di regressione $\beta = \frac{\sigma_{XY}}{\sigma_X^2}$
 - Nel seguito si userà

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

$$S_{XY} = \frac{N}{N-1} \sigma_{XY}$$

Notazioni sul campione

- s campione
- n dimensione del campione

$$\begin{array}{l} \text{osservazioni} \\ \text{valori di } y \end{array} \begin{bmatrix} 1 & 2 & \cdot & \cdot & n \\ y_1 & y_2 & \cdot & \cdot & y_n \end{bmatrix}$$

- Media campionaria $\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$
- Varianza campionaria $s^2 = \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2$

Notazioni sul campione

- Covarianza campionaria $s_{xy} = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$
- Coefficiente di correlazione campionario $r_{xy} = \frac{s_{xy}}{s_x s_y}$
- Coefficiente di regressione campionario $b = \frac{s_{xy}}{s_x^2}$