Elementi di Statistica

Contenuti

Contenuti di Statistica nel corso di Data Base

- ✓ Elementi di statistica descrittiva: media, moda, mediana, indici di dispersione
- ✓ Introduzione alle variabili casuali e alle distribuzioni di probabilità: caratteristiche di una variabile casuale, principali distribuzioni di probabilità, momenti di una distribuzione di probabilità
- ✓ Elementi di statistica induttiva: verifica di un'ipotesi, campionamento, stima di una variabile incognita

Testo di riferimento: S. Draghici, "Data Analysis Tools for DNA Microarrays", Chapman & Hall, 2003

+ Dispense

Introduzione

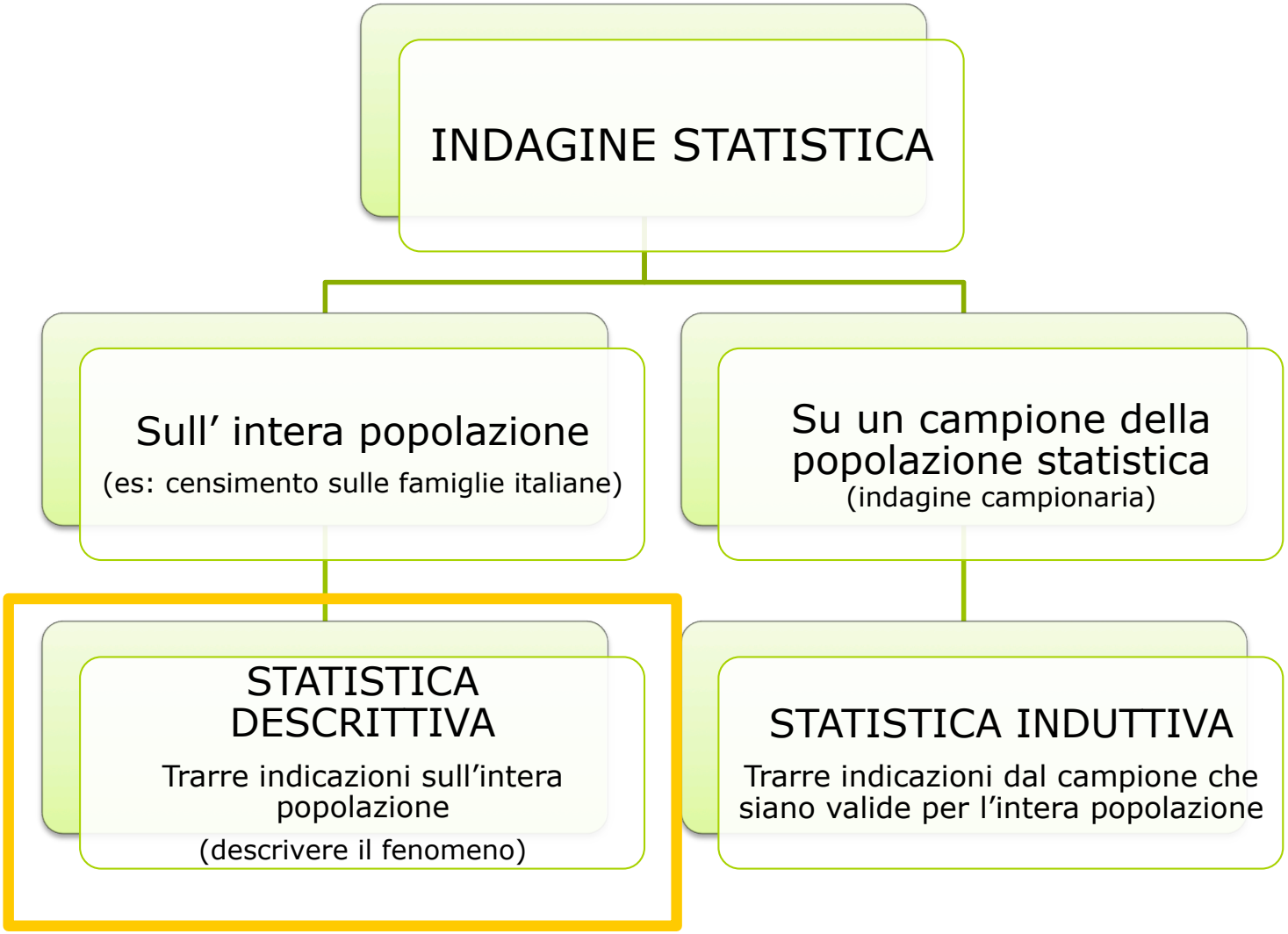
- Statistica: raccolta di metodi e strumenti matematici atti ad organizzare una o più serie di dati che descrivono una categoria di fatti

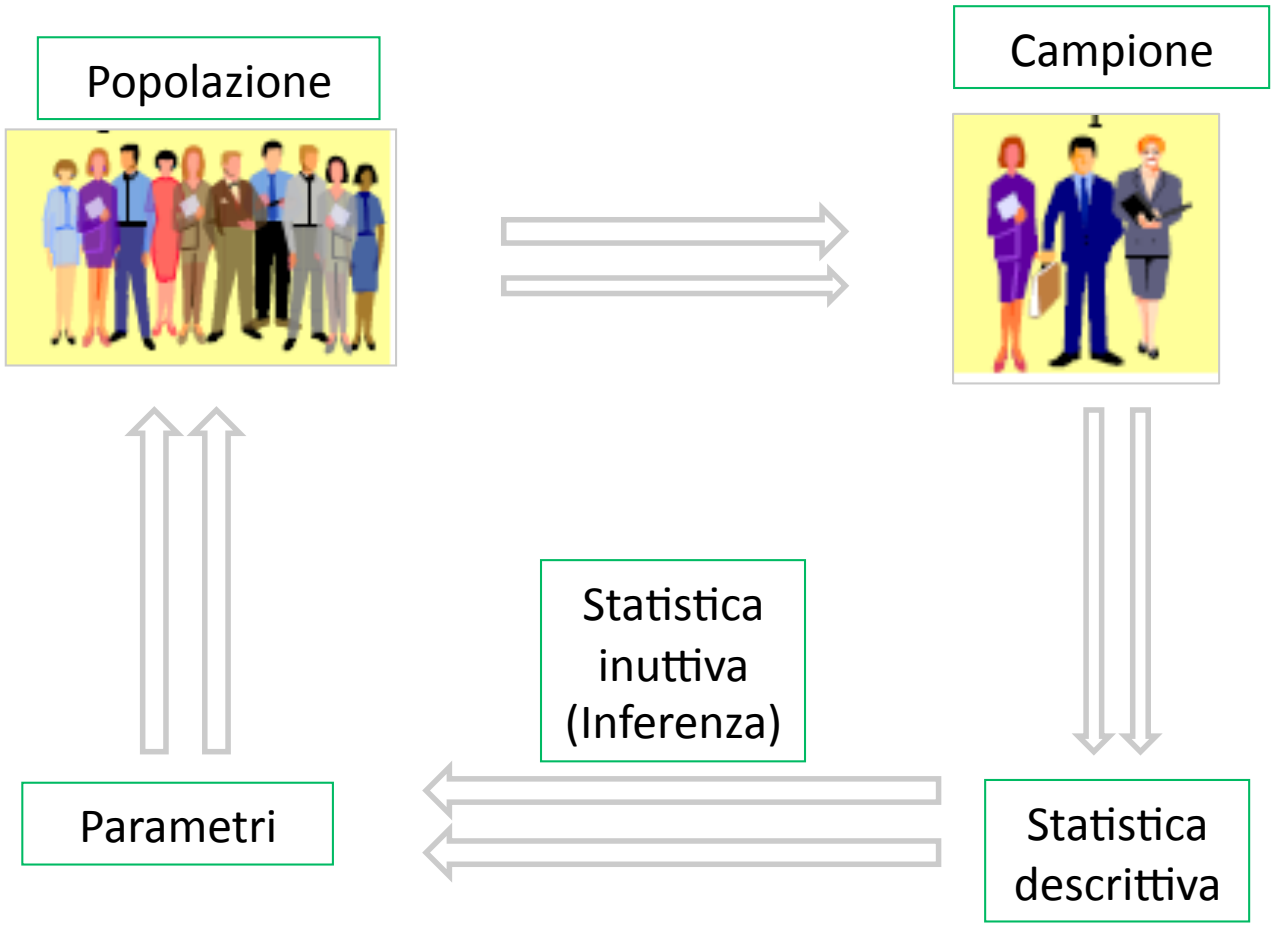
- È la scienza che studia i *fenomeni collettivi o di massa*.

  - Esempi: numero di componenti delle famiglie di una data area geografica, l'età dei cittadini di un certo paese, la lunghezza delle foglie di un tipo di pianta,la durata delle lampadine di una certa marca,…

- La statistica insegna a individuare i modi in cui un fenomeno si manifesta, a descriverlo sinteticamente, e a trarne da esso conclusioni più generali di fenomeni più ampi.

Indagine statistica




Popolazione
Campione
Statistica inuttiva (Inferenza)
Parametri
Statistica descrittiva

Popolazione, unità, campione statistico

- **Popolazione statistica:** insieme degli elementi a cui si riferisce l'indagine statistica:
  - Esempi:
    - opinione degli americani riguardo una nuova elezione presidenziale: tutti i cittadini USA.
    - geni sovra-espressi nelle persone che soffrono di obesità: tutte le persone obese
    - …
- **Unità statistica:** ogni elemento della popolazione statistica, la minima unità della quale si raccolgono i dati:
  - Un cittadino, una persona obesa….
- **Campione statistico (*sample*):** un qualsiasi insieme di unità statistiche prese da tutta la popolazione. Un campione è dunque un sottoinsieme di misurazioni selezionate dalla popolazione
  - Esempio: 50 persone con problemi di obesità (estratte a caso).

Variabile casuale

- Il *fenomeno collettivo* si presenta secondo modalità diverse nelle varie unità statistiche, perciò lo chiameremo variabile casuale.

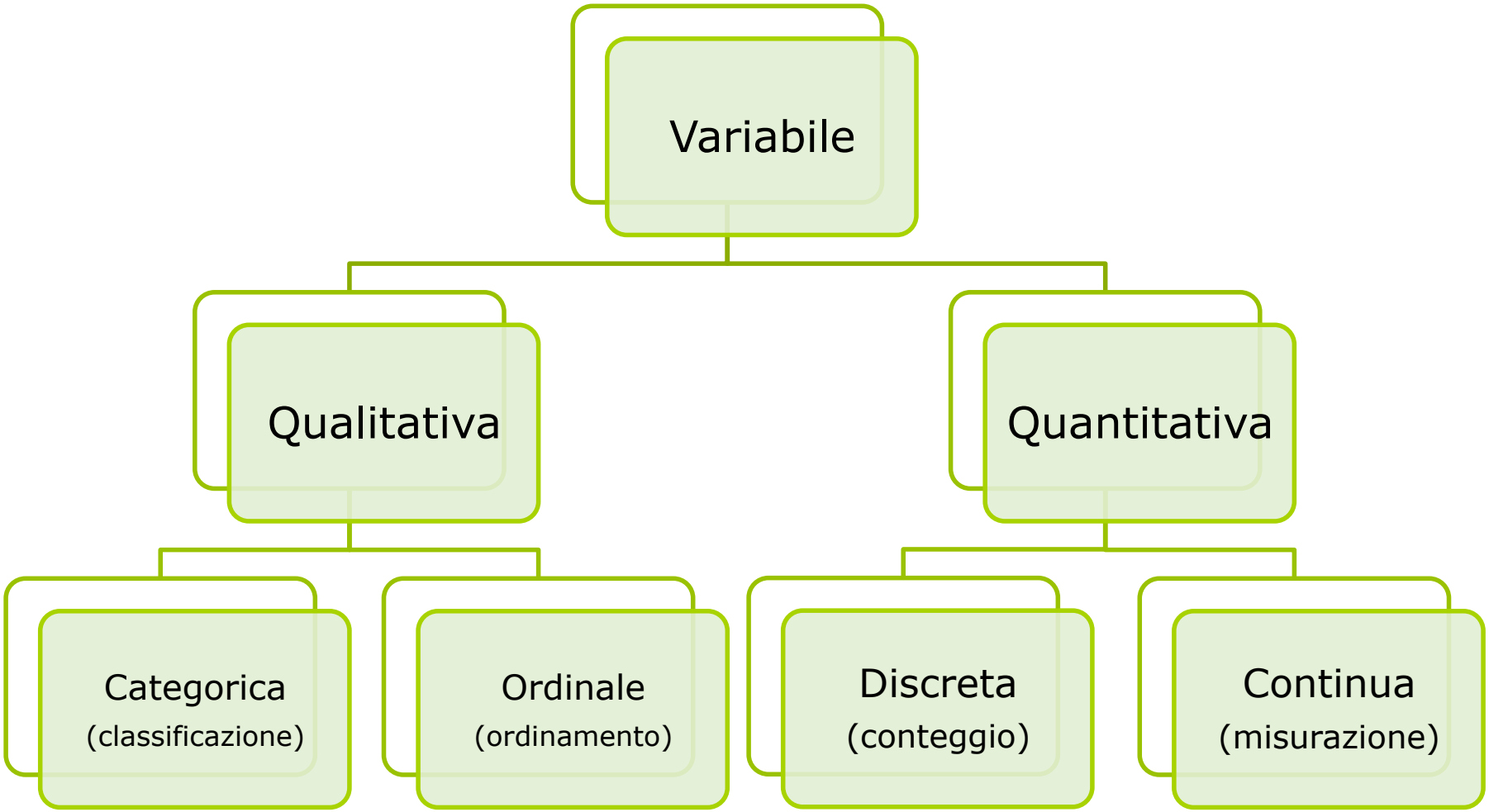
- Il valore assunto dalla variabile casuale in una data unità statistica lo chiameremo **osservazione.**

- *Esempio:*
  - *variabile casuale:* livello di espressione del gene AAA;
  - *osservazione*: il gene *AAA della persona X* ha un livello di espressione pari a 12.3, il gene AAA della persona *Y* ha un livello di espressione di 10.2, il gene *AAA della persona Z*….

Variabile quantitativa e qualitativa

- **Variabile quantitativa:** quando assume valori numerici:
  - ***Continua***: assume valori continui in un intervallo (peso e statura di una persona, livelli di intensità dei campioni su microarray, livello di espressione genica, etc.)
  - ***Discreta***: assume valori discreti come numero di campioni, numero di geni sovra-espresso, numero di pazienti, etc.
- **Variabile qualitativa:** quando assume valori non numerici
  - ***Ordinale***: i dati sono in un ordine, come ad esempio la top ten degli artisti musicali
  - ***Categorica***: uomo/donna, basso/medio/alto, fenotipo, gruppi di pazienti malati/sani, etc.

Variabile casuale



La matrice dei dati

- I dati codificati di una rilevazione statistica effettuata su *n unità statistiche* con riferimento a *p variabili*, vengono raccolti in una tabella che viene chiamata "*matrice dei dati*"

N.	Sesso	Titolo di studio	Età	Peso	N. Ricoveri
1	M	Licenza media inferiore	36	65	3
2	F	Laurea	45	70	1
…	…	…	…	…	…
N	F	Diploma	60	55	6

La matrice dei dati

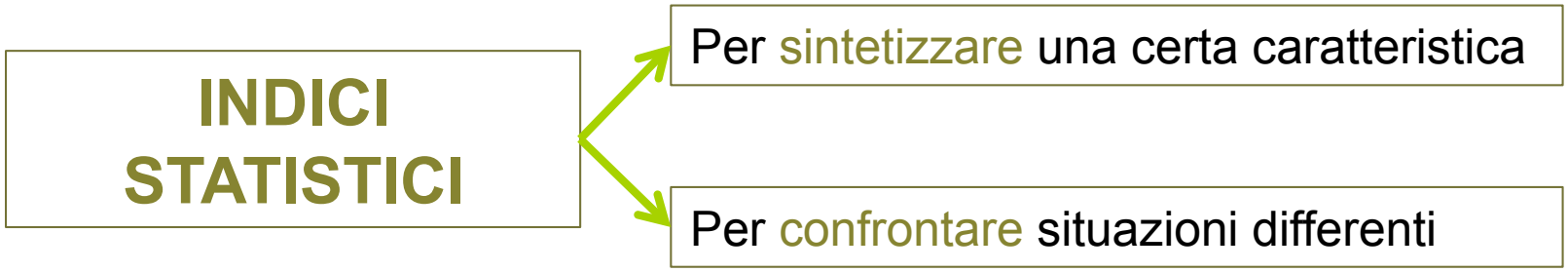
Ogni riga rappresenta un'*unità statistica*

Ogni colonna rappresenta una *variabile*

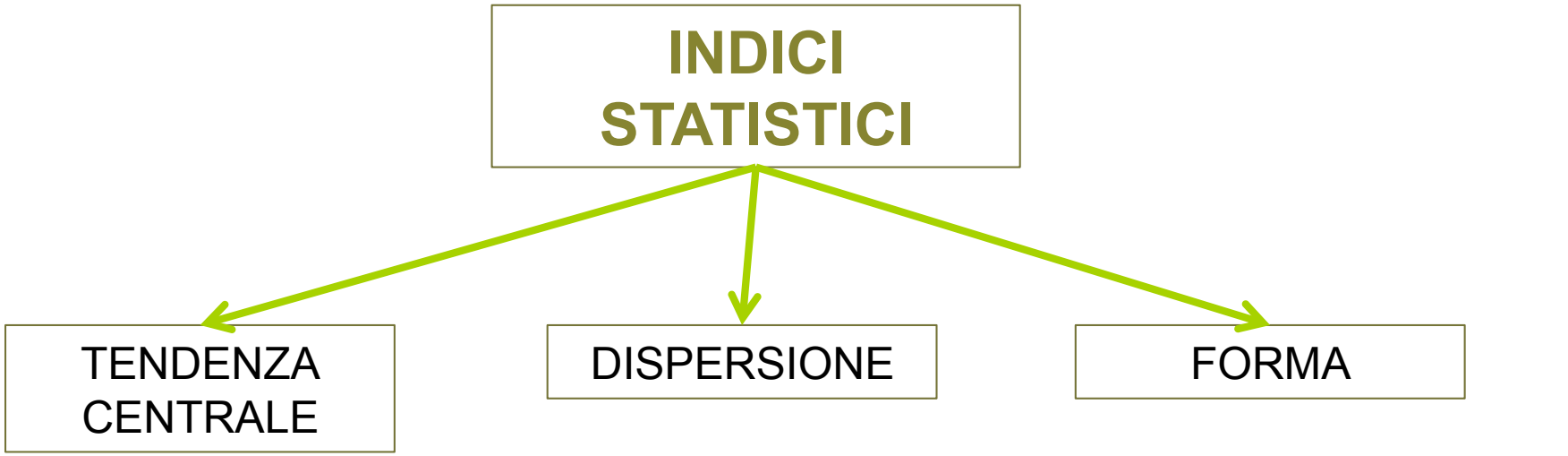
N.	Sesso	Titolo di studio	Età	Peso	N. Ricoveri
1	M	Licenza media inferiore	36	65	3
2	F	Laurea	45	70	1
…	…	…	…	…	…
N	F	Diploma	60	55	6

Analisi dei dati

- La matrice dei dati contiene tutte le informazioni analitiche di ciascuna unità statistica
- Quando i dati sono molti, l'analisi diretta della matrice non consente di cogliere in via immediata gli aspetti salienti del fenomeno
- Occorre perciò ottenere una sintesi attraverso un'elaborazione statistica dei dati



Indici statistici

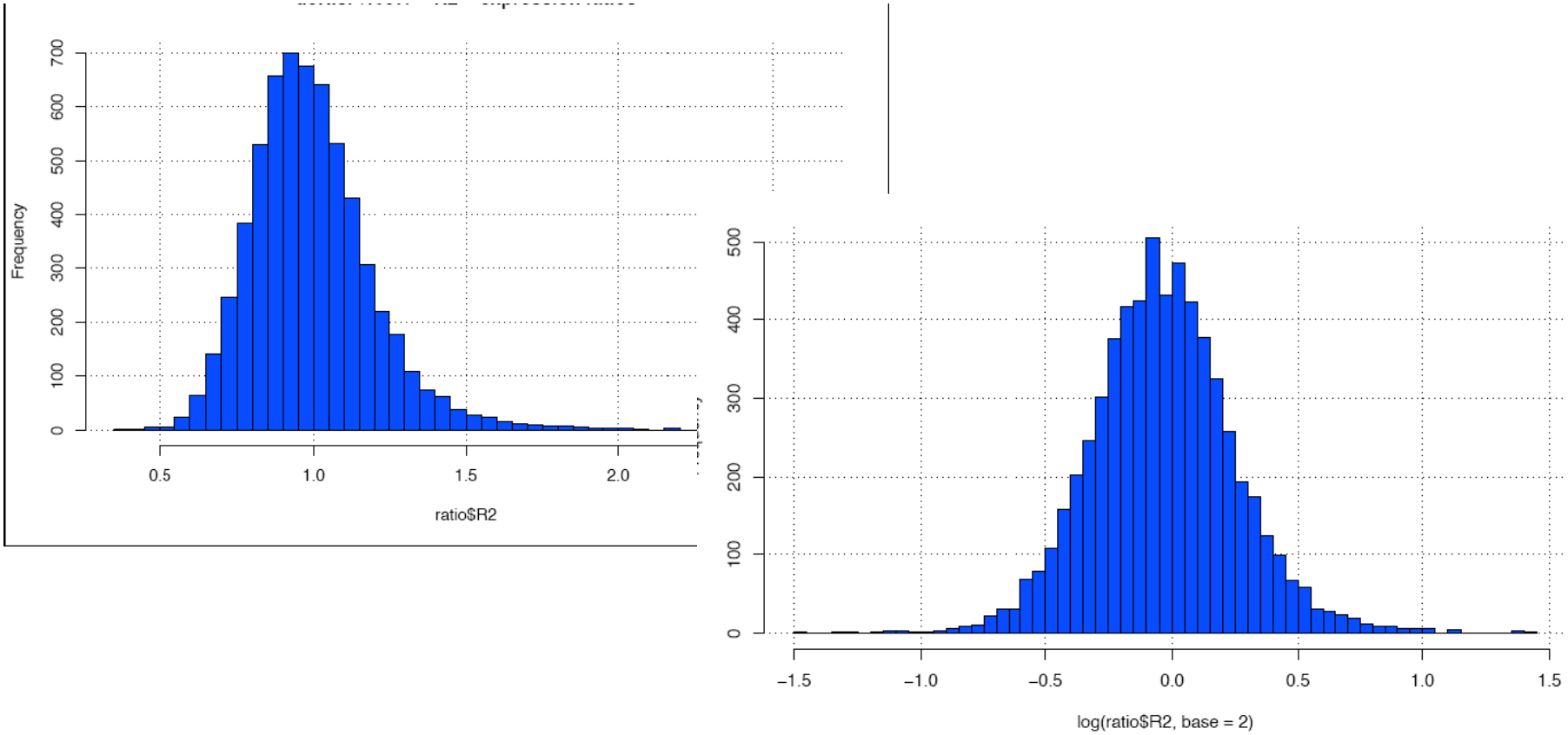


- Media
- Mediana
- Moda
- Quantili
- Percentili

- Campo di variazione
- Scarto medio assoluto
- Varianza
- Deviazione standard
- Coefficiente di variazione

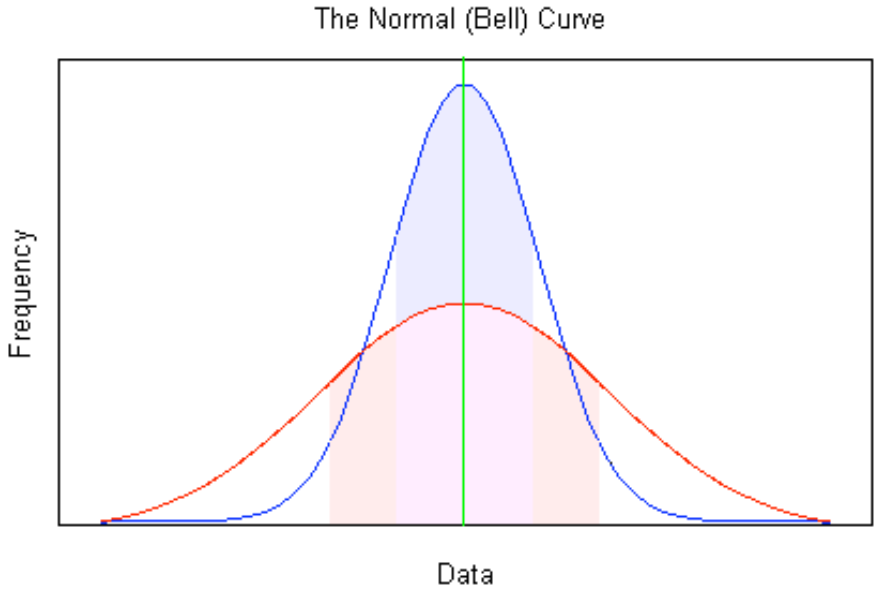
- Coefficiente di asimmetria
- Coefficiente di curtosi

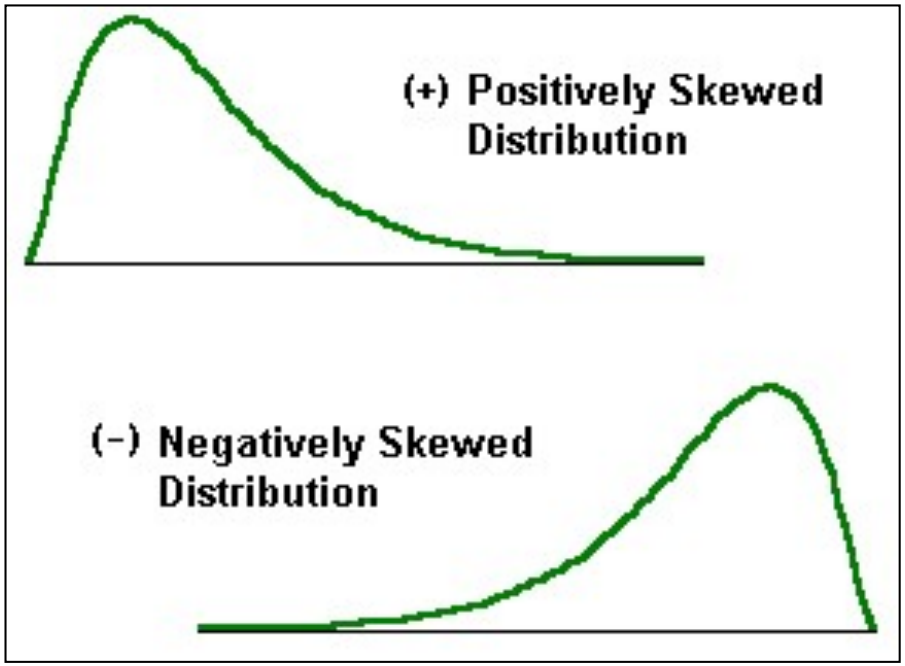
Istogramma



l'area della porzione di istogramma compresa nell'intervallo (*a*, *b*) è uguale alla frequenza relativa dei dati compresi tra *a* e *b*

Esempi


The Normal (Bell) Curve
Frequency
Data


(+) Positively Skewed Distribution
(-) Negatively Skewed Distribution

Indici di tendenza centrale

- Un indice di tendenza centrale è lo scalare che esprime sinteticamente come si è manifestata la proprietà in esame nel campione considerato.
- Può essere visto come il valore che meglio rappresenta una distribuzione: ad esempio il valore più frequente, oppure il valore che occupa una posizione intermedia nella distribuzione.

- Indici analizzati:
  - MEDIA
  - MODA
  - MEDIANA
  - QUANTILI

Media

- Media di una popolazione: somma di tutti i valori delle variabili della popolazione diviso il numero di unità della popolazione (*N)*

$$\mu = \frac{\sum_{i=1}^{N} X_i}{N}$$

Dove:
-N = numero elementi popolazione
-*Xi* =i-esima osservazione della variabile Xi

- Media di un campione: somma di tutti i valori delle variabili di un sottoinsieme della popolazione diviso il numero di unità di tale campione (*n)*

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

Media - esempio

Dato il seguente set di misurazioni di livello di espressione dei geni:

55.20	18.06	28.16	44.14	61.61	4.88	180.29	399.11	97.47	56.89	271.95	365.29	807.80

Media della popolazione:

$$\mu = \frac{\sum_{i=1}^{13} 55.20 + 18.06 + 28.16 + 44.14 + 61.61 + \ldots + \ldots + 807.80}{13} = \frac{2390{,}85}{13} = 183.9115$$

Media del campione (55.20; 18.06; 28.16; 44.14):

$$\overline{X} = \frac{55.20 + 18.06 + 28.16 + 44.14}{4} = \frac{145.56}{4} = 36.39$$

La media di qualsiasi campione $\overline{X}$ può essere molto diversa da quella dell'intera popolazione μ.
Più è numeroso il campione, più la media del campione sarà vicina a quella della popolazione.

Valore atteso e campionamento

- Il ***valore atteso*** di una variabile X, indicato con E[X] è definito come la media di X calcolata su un grande numero di esperimenti

Campionamento con rimpiazzo e senza rimpiazzo:

- Se un campione è costruito prendendo un valore e successivamente eliminando quel valore dalla popolazione in modo tale che non possa essere preso nuovamente, si dice che il ***campionamento*** è effettuato ***senza rimpiazzo***

- Se il valore usato in un campione non è rimosso dalla popolazione in modo tale che lo stesso valore possa essere preso nuovamente, si dice che il ***campionamento*** è effettuato ***con rimpiazzo***

Media

- Media ponderata di una popolazione: si assegna ad ogni variabile un peso; si sommano tutti i valori delle variabili, moltiplicate per il peso, e si divide il numero ottenuto per la somma dei pesi

$$\mu = \frac{\sum_{i=1}^{N} p_i X_i}{\sum_{i=1}^{N} p_i}$$

Esempio: calcolo media voti

Moda

- La moda è il valore più frequente di una distribuzione, o meglio, la modalità più ricorrente della variabile (cioè quelle a cui corrisponde la frequenza più elevata).

962	1005	1003	768	980	965	1030	1005	975	989	955	783	1005

La moda di questo campione è 1005 in quanto compare ben 3 volte.

- Caratteristiche:
  - viene utilizzata solamente a scopi descrittivi, perché **è meno stabile e meno oggettiva delle altre misure di tendenza centrale.**
  - Per individuare la moda di una distribuzione si possono usare gli istogrammi,
  - Può differire nella stessa serie di dati, quando si formano classi di distribuzione (intervalli) con ampiezza differente.
  - Per individuare la moda entro una classe di frequenza, non conoscendo come i dati sono distribuiti, si ricorre all'ipotesi della ripartizione uniforme.

Distribuzioni unimodali/bimodali

Una distribuzione può presentare più mode:

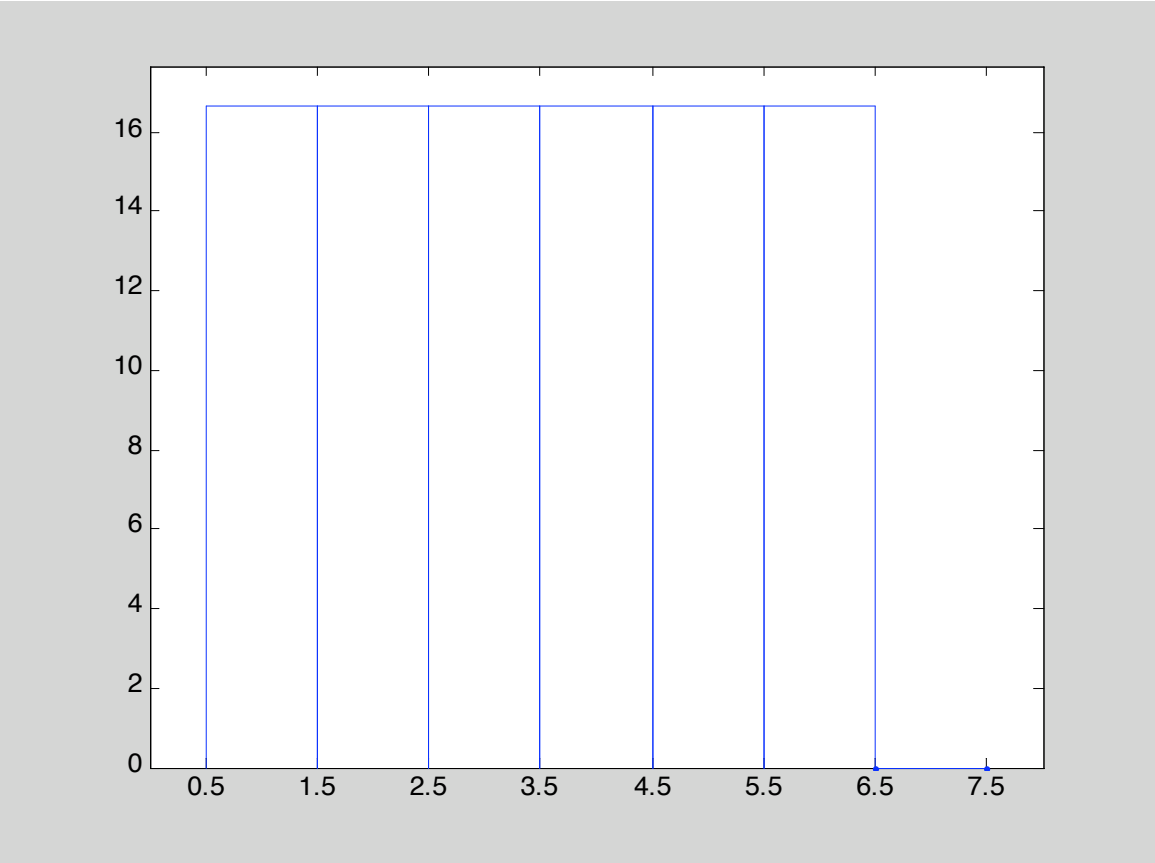
- **Distribuzioni unimodali:** distribuzioni di frequenza che hanno una sola moda, ossia un solo un punto di massimo (che rappresenta sia il massimo relativo che il massimo assoluto);

- **Distribuzioni bimodali o k-modali**: distribuzioni di frequenza che presentano due o più mode, ossia che hanno due (o k) massimi relativi;
  - *Esempio:* misurando le altezze di un gruppo di giovani in cui la parte maggiore sia formata da femmine e la minore da maschi si ottiene una distribuzione bimodale, con una moda principale ed una secondaria.

Distribuzione zeromodale

Nessun valore ha una frequenza più elevata degli altri.

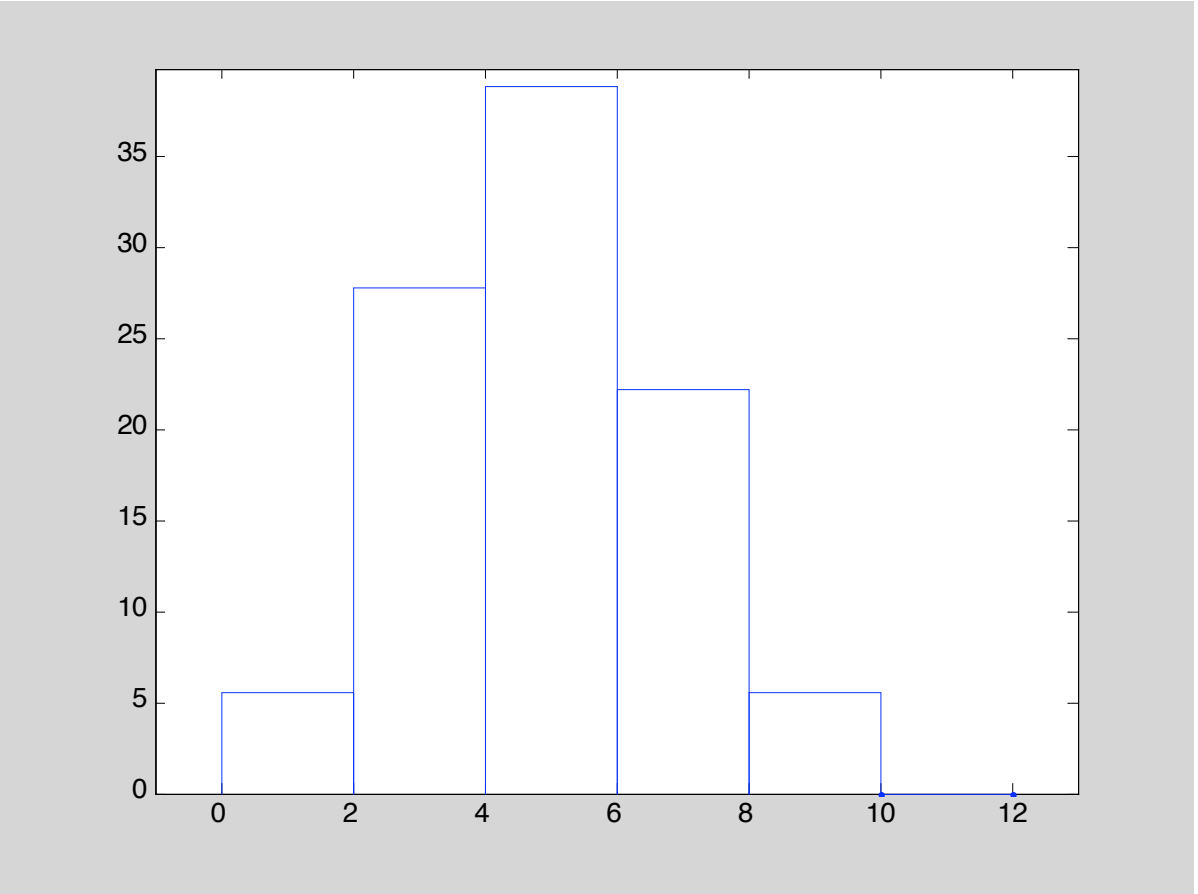
A = {1, 1, 1, 2, 2, 2, 3, 3, 3, 4, 4, 4, 5, 5, 5, 6, 6, 6}



Distribuzione unimodale

C’è un solo valore con una frequenza più elevata degli altri

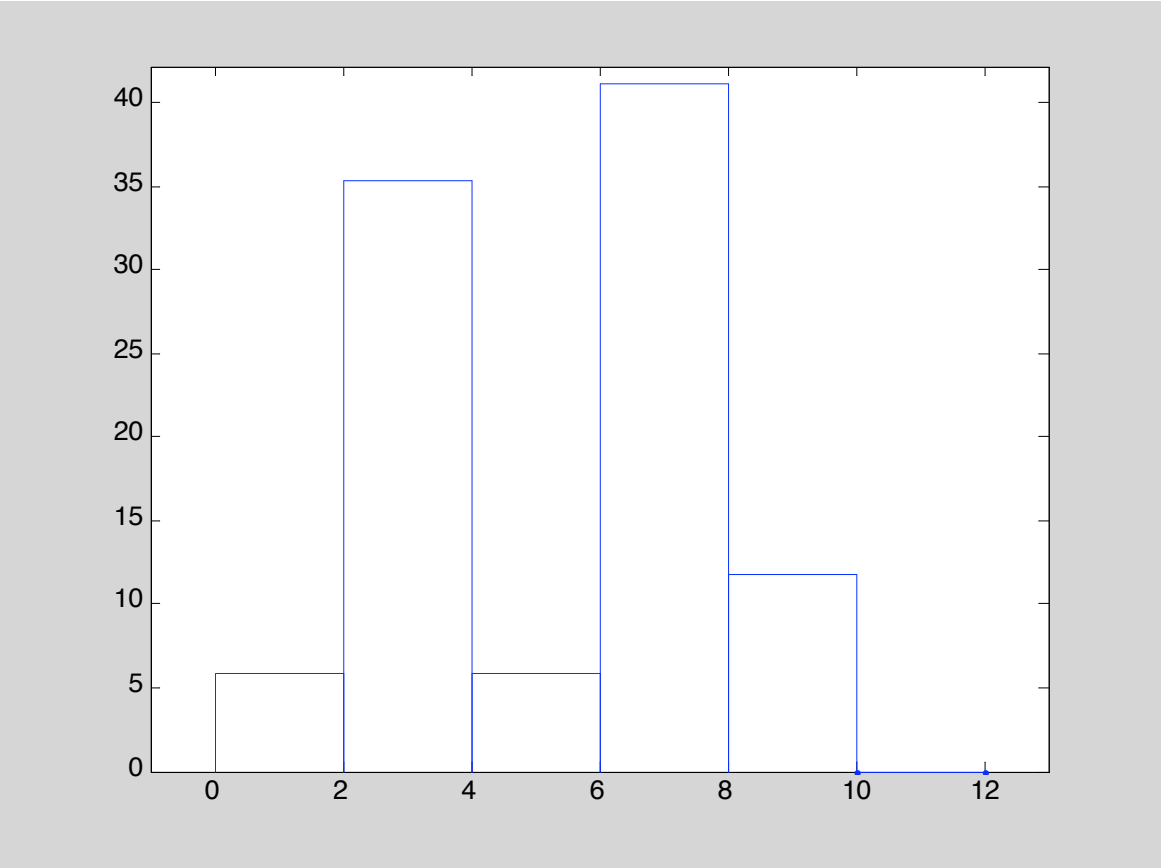
A = {1, 2, 2, 3, 3, 3, 4, 4, 4, 4, 5, 5, 5, 6, 6, 7, 7, 8}



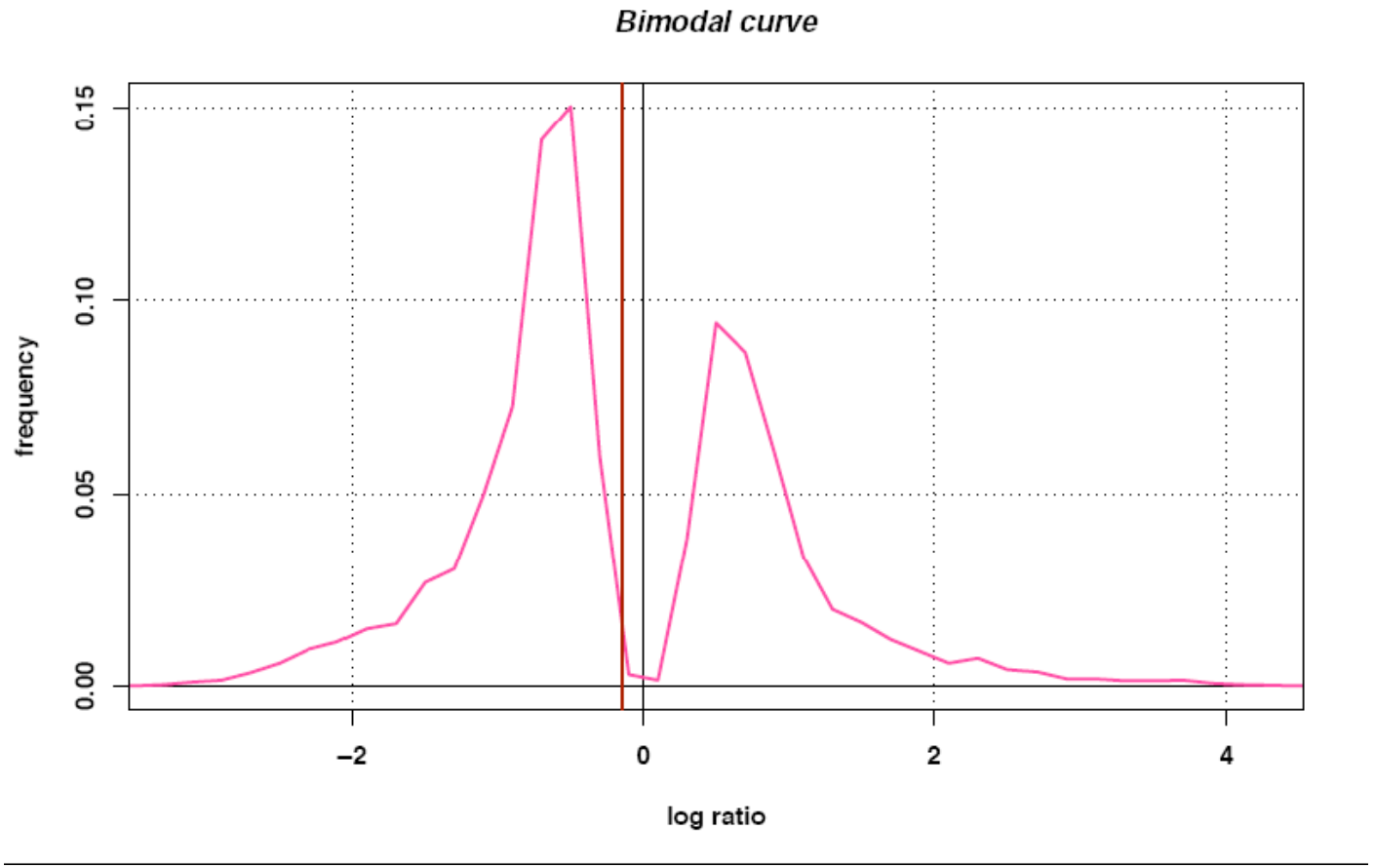
Distribuzione bimodale

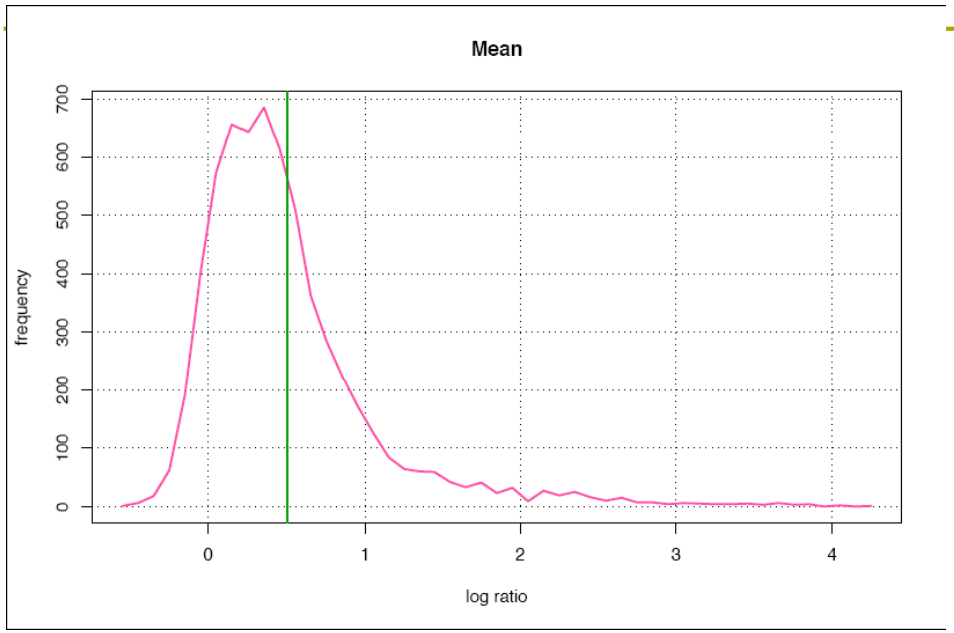
Ci sono due valori con una frequenza più elevata degli altri.

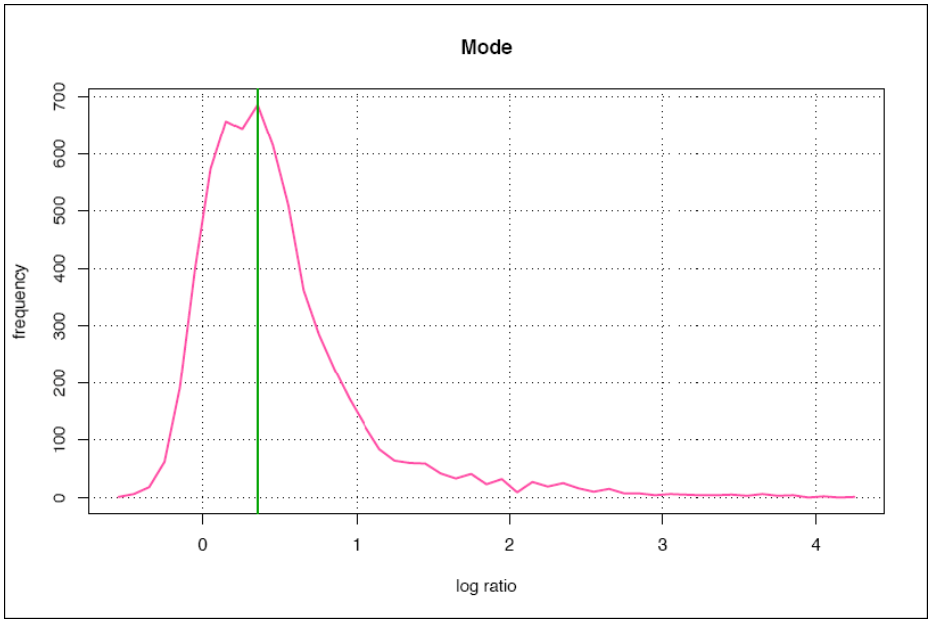
A = {1, 2, 2, 3, 3, 3, 3, 5, 6, 6, 6, 6, 6, 7, 7, 8, 8}



Distribuzione Bimodale




Mean
frequency
0
100
200
300
400
500
600
700
0
1
2
3
4
log ratio


Mode
frequency
0
100
200
300
400
500
600
700
0
1
2
3
4
log ratio

Mediana

- La mediana è il valore che occupa la posizione centrale in un insieme ordinato di dati.
- E’ una misura robusta, in quanto poco influenzata dalla presenza di dati anomali.
- Caratteristiche:
  - si ricorre al suo uso quando si vuole attenuare l'effetto di valori estremi;
  - in una distribuzione o serie di dati, ogni valore estratto a caso ha la stessa probabilità di essere inferiore o superiore alla mediana.

Mediana- calcolo

Per calcolare la mediana di un gruppo di dati, bisogna:

1.disporre i valori in ordine crescente oppure decrescente e contare il numero totale *n* di dati;

2.se il numero (*n*) di dati è dispari, la mediana corrisponde al valore numerico del dato centrale, quello che occupa la posizione *(n+1)/2*;

3.se il numero (*n*) di dati è pari, la mediana è stimata utilizzando i due valori centrali che occupano le posizioni *n/2* e *n/2+1*:

a. con poche osservazioni, come mediana viene assunta la media aritmetica di queste due osservazioni intermedie;
b. con molte osservazioni raggruppate in classi, si ricorre talvolta alle proporzioni.

Mediana-esempio

Consideriamo il seguente campione:

96	78	90	62	73	89	92	84	76	86

1. Ordiniamo i campioni in ordine crescente:

62	73	76	78	**84**	**86**	89	90	92	95

2. Dal momento che il numero di campioni è pari *(n=10)* la mediana è calcolata come la media dei due elementi centrali*:*

$$mediana = \frac{84 + 86}{2} = 85$$

Esempio

Voti di Pierino primo quadrimestre

1	5	5	5	6	6	6

Moda ? Mediana ? Media ?

Voti di Pierino secondo quadrimestre

4	5	5	5	6	7	10

Moda ? Mediana ? Media ?

Quantili

- I quantili sono una famiglia di misure, a cui appartiene anche la mediana, che si distinguono a seconda del numero di parti uguali in cui suddividono una distribuzione.

- I quartili ripartiscono la distribuzione in 4 parti di pari frequenza, dove ogni parte contiene la stessa frazione di osservazioni:
  - Il **primo quartile** è definito come il numero *q1* per il quale il *25%* dei dati statistici è minore o uguale a *q1.*
  - Il **secondo quartile** è definito come il numero *q2* per il quale il *50%* dei dati statistici è minore o uguale a *q2*. Il secondo quartile corrisponde alla mediana
  - Il **terzo quartile** è definito come un numero *q3* per il quale il *75%* dei dati statistici è minore o uguale a *q3.*

Quartili- esempio

Studio che esamina i tempi d'attesa al ristorante in un campione di 10 clienti

Dati ordinati:

Q2 = Mediana

58.6 59.0 59,3 59,4 62,7 62,8 63,7 65,4 67,3 68,1

La mediana è pari a 62,75
Si considera la metà inferiore dei dati, ovvero tutti i valori inferiori alla mediana e su questo sottoinsieme di dati si calcola la mediana, il valore trovato è Q1

Q1

58.6 59.0 59,3 59,4 62,7

Si considera la metà superiore dei dati, ovvero tutti i valori superiori della mediana e su questo sottoinsieme di dati si calcola la mediana il valore trovato è Q3

62,8 63,7 65,4 67,3 68,1

Q3

Decili e percentili

- In modo analogo si definiscono i:
  - Decili: 9 punti che dividono la distribuzione ordinata in 10 parti uguali
  - Percentili: 99 punti che dividono la distribuzione ordinata in 100 parti uguali.

Indici di dispersione

- Un indice di dispersione restituisce uno scalare con cui si valuta la diversità esistente tra le osservazioni.

- Indici analizzati:
  - CAMPO DI VARIAZIONE
  - VARIANZA
  - DEVIAZIONE STANDARD
  - COVARIANZA
  - CORRELAZIONE

Campo di variazione (o "range")

- Il campo di variazione di una distribuzione è la differenza tra il dato più grande e quello più piccolo della distribuzione:

$$C = x_{max} - x_{min}$$

- Questo indice è abbastanza grossolano non dicendo nulla sulla variabilità dei dati intermedi.
- *Esempio:* il campo di variazione della seguente distribuzione: 25 – 26 – 28 – 29 – 30 – 32
  è *C*= 32 – 25= 7

Scarto

- Lo **scarto** misura quanto ciascun dato x_i si discosta dal valor medio, ovvero s= $x_i - \overline{X}$

Esempio:

Consideriamo le seguenti intensità rilevate dagli spot dei microarray:

435.02, 678.14, 235.35, 956.12, …, 1127.82, 456.43

La media di questi valori è: 515.13;

i loro scarti sono:

435.02 - 515.13 = -80.11
678.14 - 515.13 = 163.01
235.35 – 515.13 = -279.78
956.12 – 515.13 = 440.99
…

Quale sarà la media di questi valori ??

Scarto assoluto

- Usando **s** possono essere ricavati diversi altri indici di variabilità

- Si chiama **scarto medio assoluto** e si indica con s_m la media aritmetica dei valori assoluti degli scarti

$$S_m = \frac{\sum_{i=1}^{n} | x_i - \overline{X} |}{n}$$

Varianza

- Varianza della popolazione: misura che caratterizza molto bene la variabilità di una popolazione

$$\sigma^2 = \frac{\sum_{i=1}^{N} (X_i - \mu)^2}{N}$$

Dove:
- N è il numero di osservazioni dell'intera popolazione
- μ è la media della popolazione
- x_i è l'i-esimo dato statistico osservato

- Varianza di un campione:

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

Dove:
- *n* è il numero di osservazioni del campione
- $\overline{X}$ è la media del campione
- x_i è l'i-esimo dato statistico osservato

- Quando *n* è grande le differenze tra le due formule sono minime; quando *n* è piccolo, le differenze sono sensibili.

Varianza - esempio

Consideriamo il seguente campione di osservazioni: {2, 3, 6, 9, 15}

Calcolo della media: $$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{2+3+6+9+15}{5} = 7$$

Calcolo della varianza campionaria: $$s^2 = \frac{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2}{n-1}$$

$$= \frac{(2-7)^2 + (3-7)^2 + (6-7)^2 + (9-7)^2 + (15-7)^2}{4} = 27{,}5$$

Devianza

Nel calcolo di alcune statistiche si ricorre alla devianza, data dal numeratore della varianza:

$$Dev = \sum_{i=1}^{N} \left(X_i - \overline{X} \right)^2$$

Deviazione standard o scarto quadratico medio

- La varianza ha lo svantaggio di essere una grandezza quadratica e quindi non direttamente confrontabile con la media o con gli altri valori della distribuzione.
- Per trovare una misura espressa nella stessa unità di misura della variabile di partenza è sufficiente estrarre la radice quadrata della varianza.
- La deviazione standard è una misura di distanza dalla media e quindi ha sempre un valore positivo.
- E' una misura della dispersione della variabile casuale intorno alla media.

Deviazione standard o scarto quadratico medio

- Deviazione standard della popolazione

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N} (X_i - \mu)^2}{N}}$$

Dove:
- N è il numero di osservazioni dell'intera popolazione
- μ è la media della popolazione
- X_i è l'i-esimo dato statistico osservato

- Deviazione standard di un campione

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

Dove:
- *n* è il numero di osservazioni del campione
- $\overline{X}$ è la media del campione
- X_i è l'i-esimo dato statistico osservato

Alcune considerazioni

- Cambiando strumento di misura posso ottenere misure tutte aumentate di una costante c

$$\overline{X}_c = \overline{X} + c$$

- La varianza rimane invariata $s_c^2 = s^2$

- Se le misure vengono moltiplicate per una costante

$$\overline{X}_c = c\overline{X}$$

$$s_c^2 = c^2 s^2$$

Deviazione standard - esempio

- Consideriamo i voti di due studenti:
  Anna (30, 30, 28, 27, 26)
  Stefano (21, 30, 30, 30, 30).

- Entrambi hanno la stessa media dei voti (media=28.2)

- Calcoliamo la deviazione standard:
  σ(Anna) = 1.78
  σ(Stefano) = 4,02

- Cosa significa?
  Significa che i voti di Anna sono più concentrati (vicini) rispetto a quelli di Stefano

Covarianza

- Indice che consente di verificare se fra due variabili statistiche esiste un legame lineare

- Considerando due serie $\{x_i\}$ e $\{y_i\}$, i=1,2,…n, pone a confronto le coppie di scarti $(x_i - \bar{x})$ e $(y_i - \bar{y})$:

$$Cov(X,Y) = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

Covarianza

La Covarianza può essere:

- *POSITIVA*: quando X e Y variano tendenzialmente nella stessa direzione, cioè al crescere della X tende a crescere anche Y e al diminuire della X tende a diminuire anche Y.

- *NEGATIVA*: quando le due variabili variano tendenzialmente in direzione opposta, cioè quando al crescere di una variabile l'altra variabile tende a diminuire (e viceversa).

- *NULLA*: quando non vi è alcuna tendenza delle 2 variabili a variare nella stessa direzione o in direzione opposta. Quando Cov(X,Y) = 0 si dice anche che X ed Y sono non correlate o linearmente indipendenti.

Correlazione

- Un modalità più rigorosa che consente di studiare il ***grado di intensità*** del legame lineare tra coppie di variabili.

$$r_{xy} = \frac{Cov(X,Y)}{\sqrt{(VarX)(VarY)}}$$

- Coefficiente di Pearson

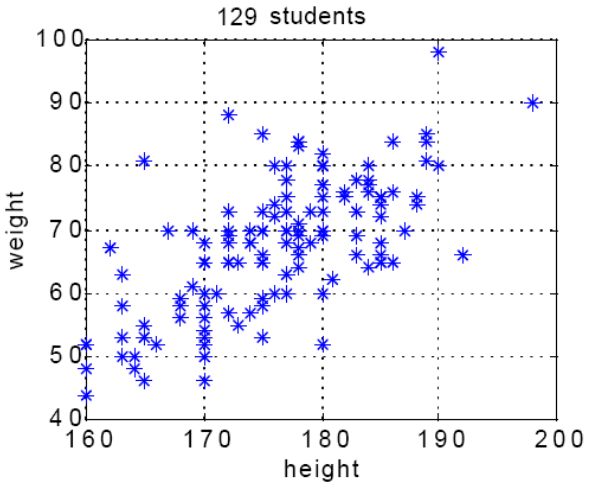
Il coefficiente di correlazione ci permette di:
✓ riassumere la forza della relazione ***lineare*** fra le variabili
✓ verificare l'apparente associazione fra le variabili

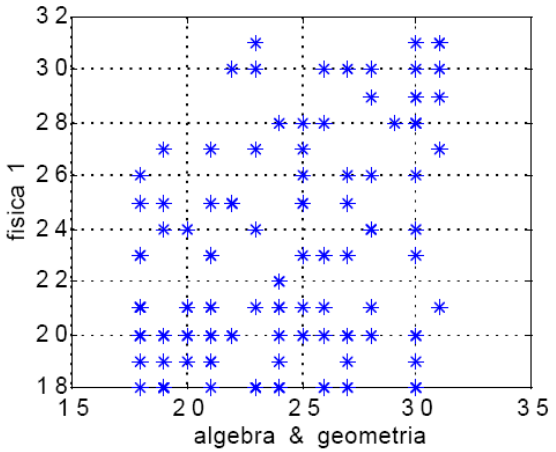
Il coefficiente di correlazione:
✓ varia da –1 a 1 (se uguale a 1 o a -1: perfettamente correlate)
✓ è positivo quando i valori delle variabili crescono insieme
✓è negativo quando i valori di una variabile crescono al decrescere dei valori dell'altra
✓ non è influenzato dalle unità di misura

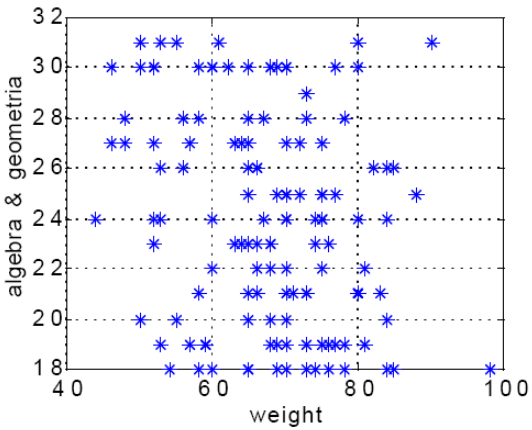
Correlazione

- Esempio:
  - Sono stati raccolti i seguenti dati da *n* = 129 studenti*:*
    - Altezza (cm)
    - Peso (Kg)
    - Voto Algebra e Geometria
    - Voto Fisica I
  - Valutare la correlazione delle seguenti coppie:
    - Peso – Altezza
    - Algebra e Geometria - Fisica I
    - Peso - Algebra e Geometria

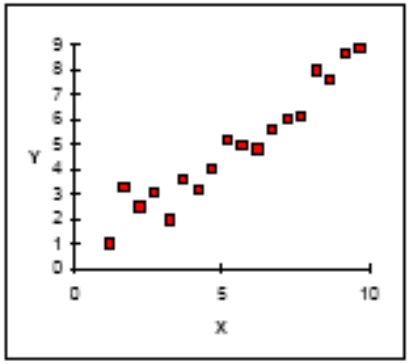
Correlazione - esempio





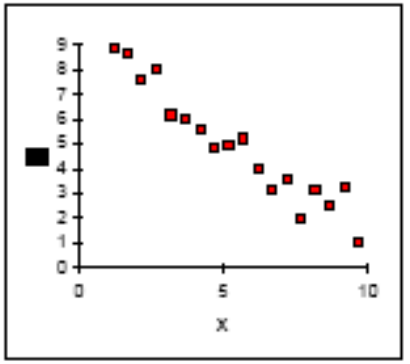


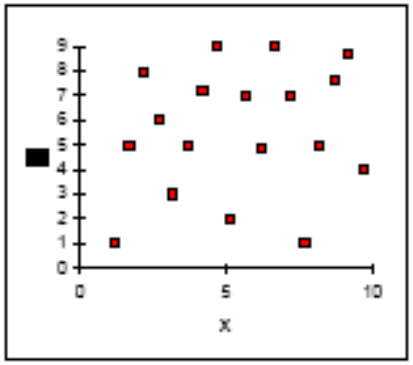
Coefficiente di correlazione - esempi



r=0,96

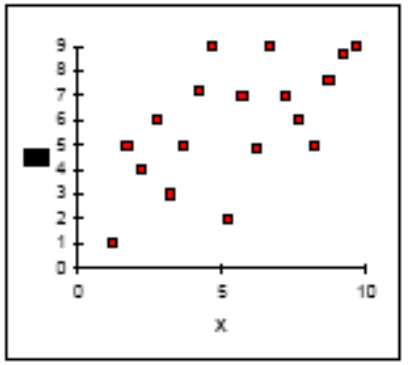
r=-0,96





r=0,12

r=0,62

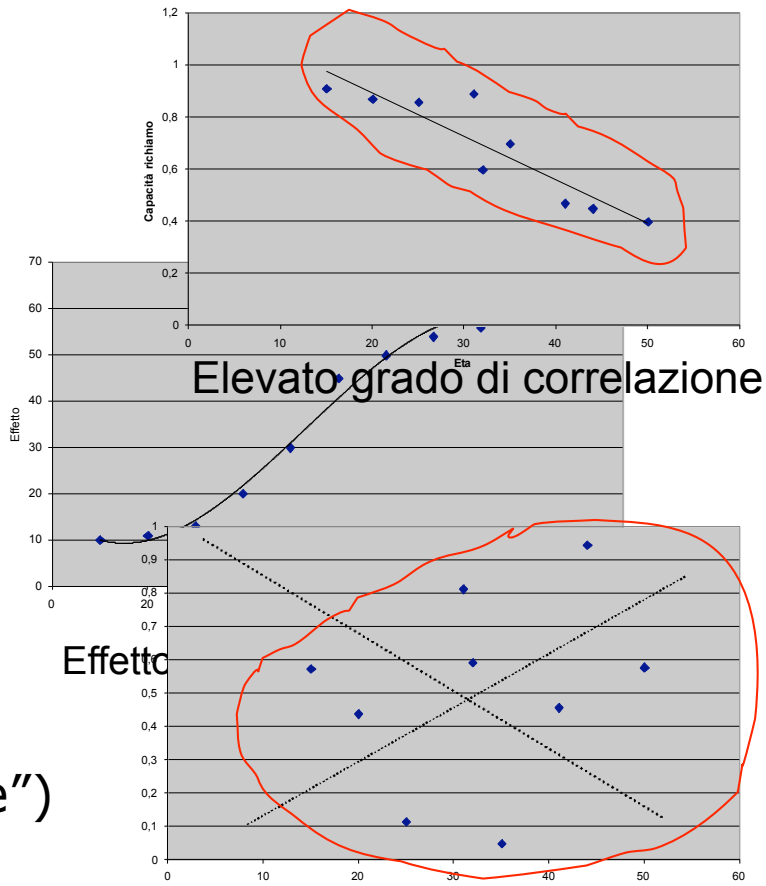


r=1: punti perfettamente allineati su una retta con pendenza positiva

r=-1: punti perfettamente allineati su una retta con pendenza negativa

Correlazione: caratteristiche

- Direzione della relazione:
  - Correlazione positiva
  - Correlazione negativa
- Forma della correlazione:
  - Lineare
  - Forme non-lineari
- Grado di correlazione:
  Le relazioni si distinguono a secondo del grado di correlazione:
  - Elevato grado di correlazione
    (punti vicini alla "linea di regressione")
  - Basso grado di correlazione
    (punti lontani dalla "linea di regressione")

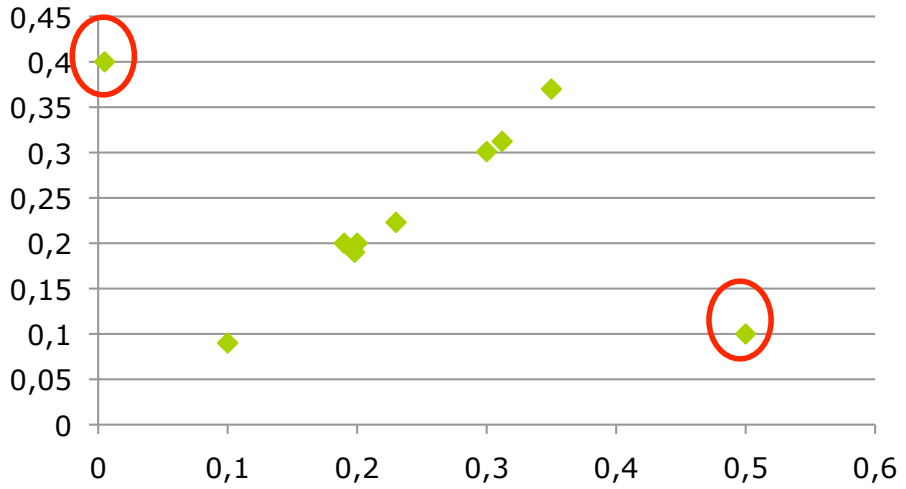


Elevato grado di correlazione

Basso grado di correlazione

Correlazione: rischi nell'interpretazione

- Un'elevata correlazione fra due variabili NON implica una relazione causa-effetto
- La correlazione non è equivalente alla dipendenza
- La correlazione è molto sensibile agli outliers: due o tre outliers possono portare il coefficiente di corrrelazione a livelli molto bassi



Matrice di covarianza

- Dato un set di variabili x_1, x_2, …, …, x_k possono essere calcolate le covarianze e le correlazioni di tutte le possibili coppie di tali variabili x_i e x_j

- Questi valori possono essere inseriti in una matrice nella quale ciascuna riga e ciascuna colonna corrisponde ad una variabile.
- L'elemento σ_{ij} situato all'intersezione tra la riga *i* e la riga *j*, sarà la covarianza tra le variabili x_i e x_j, mentre gli elementi situati sulla diagonale sarranno le varianze delle rispettive variabili:

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} & \cdots & \sigma_{1k} \\ \sigma_{21} & \sigma_2^2 & \sigma_{23} & \cdots & \vdots \\ \vdots & & & & \\ \sigma_{k1} & \sigma_{k2} & \sigma_{k3} & & \sigma_k^2 \end{bmatrix}$$

Matrice di correlazione

- La matrice di correlazione è ottenuta prendendo l'elemento ij da Σ (matrice di covarianza) e dividendolo per $\sqrt{\sigma_i^2 \sigma_j^2}$

$$r_{ij} = \frac{\sigma_{ij}}{\sqrt{\sigma_i^2 \sigma_j^2}}$$

- Le matrici di correlazione e di covarianza sono simmetriche