

Elementi di Probabilità e Statistica

Statistica Descrittiva

- Rappresentazione dei dati mediante tabelle e grafici
- Estrapolazione di indici sintetici in grado di fornire informazioni riguardo alla distribuzione dei dati, la forma, la variabilità e la tendenza centrale

Statistica Matematica

- Calcolo delle probabilità
- Variabili aleatorie e modelli teorici di distribuzione

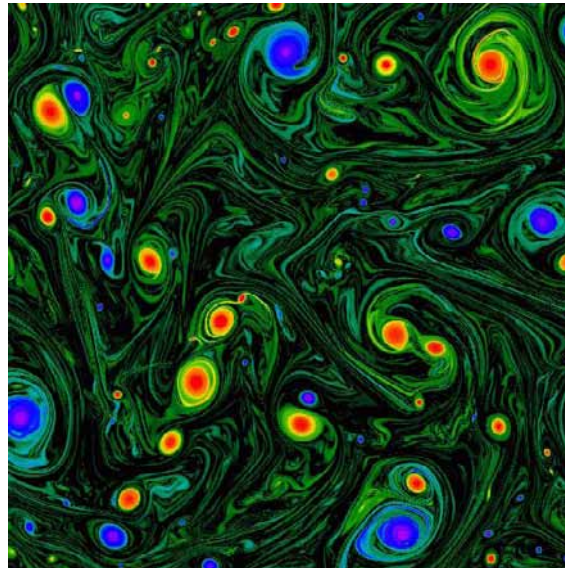
Statistica Inferenziale

- Ipotesi parametriche (su media e varianza)
- Ipotesi funzionali (su l'intera distribuzione)

Processi
Deterministici

Processi
Casuali

Chaos Deterministico?



Un po' di bibliografia ...

- *Statistics, probability and reliability for civil and environmental engineers*, N.T.Kottegoda, R. Rosso, McGraw-Hill, 1997
- *Teoria della probabilità*, E.S. Ventsel, Edizioni MIR, 1983
- *Probability, Statistics, and Decisions for Civil Engineers*, J. Benjamin, C. A. Cornell, McGraw-Hill, 1970

Dove verranno via via pubblicate le risorse didattiche del corso...

www.diam.unige.it/costid

Statistica: Scienza delle decisioni in condizioni di incertezza...

A che tipo di fenomeni si applica?

Fenomeni caratterizzati da:

- elevata variabilità (spazio-temporale)
- scarsa predicibilità
- numero elevatissimo (o infinito) di gradi di libertà

Su quali assunzioni si basa?

Sintesi delle informazioni: Il fenomeno che andiamo a descrivere tramite Le metodologie statistiche deve essere un fenomeno **collettivo**, per il quale si possano definire degli indicatori sintetici di confronto, valutazione e decisione

Quali sono le fasi fondamentali di un'analisi statistica?



Alcune definizioni di base ...

Universo (o Popolazione): Insieme oggetto del nostro studio, su cui vengono effettuate le rilevazioni statistiche. L'analisi statistica verrà condotta su suoi specifici sottoinsiemi (detti **Campioni**), rappresentativi delle caratteristiche dell'intera popolazione). Si indica in generale con la lettera ***U***

Individui: Elementi che costituiscono la popolazione.

$$U = \{u_i\}_{i=1}^N \quad \text{con } N = \text{numerosità della popolazione (quando non infinita)}$$

Gli u_i si possono anche chiamare **osservabili** o **unità**

Statistica Descrittiva —→ Ha lo scopo di individuare ed evidenziare le caratteristiche fondamentali del campione

Caratteristiche di un individuo in senso statistico

Statisticamente, una **caratteristica** non è altro che una funzione X La quale associa ad ogni individuo della popolazione un valore numerico o ordinale. X è anche detta variabile della popolazione

Spazio campionario E

E' l'insieme di tutti i valori possibili di una certa caratteristica degli individui (eventi elementari).

Può essere:

- Continuo
- Discreto

Si ha quindi in generale:

$$X : U \rightarrow E \subseteq R$$

$$X(u_i) = (x_i)$$

modalità della
variabile

Classificazione delle variabili

Variabili Qualitative

Sconnesse o
Nominali

modalità non
ordinabili

Ordinali
modalità
ordinabili

Variabili Quantitative

Discrete

L'insieme delle
modalità è finito
o numerabile

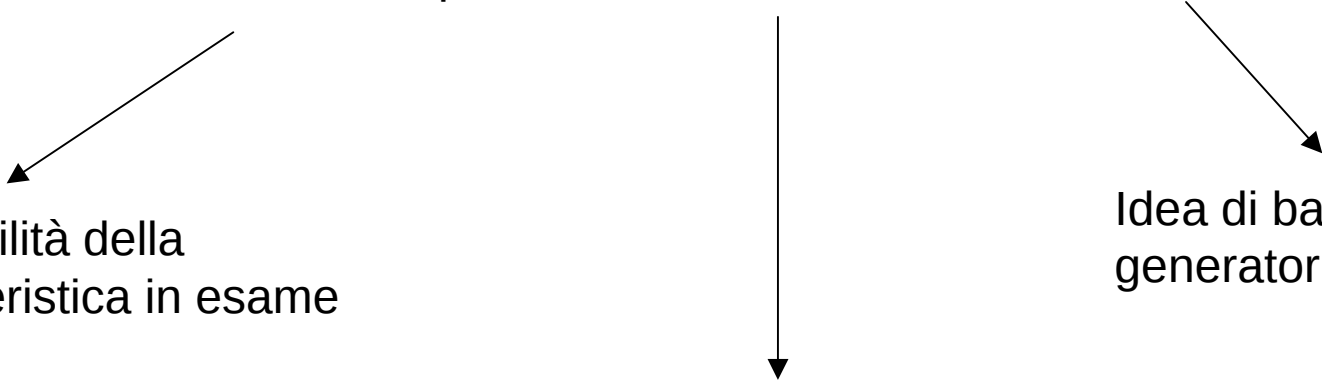
Continue

L'insieme delle
modalità è
infinito
e non
numerabile

Metodi per la rappresentazione grafica di una caratteristica quantitativa

La corretta rappresentazione grafica dei dati costituisce un passaggio fondamentale dell'analisi statistica in quanto permette di stimare in modo diretto ed intuitivo

Le caratteristiche del campione



Variabilità della
caratteristica in esame

Struttura del
set di dati

Idea di base dei processi
generatori

Metodi per la rappresentazione grafica di una caratteristica quantitativa (2)

Metodi puramente descrittivi

- Diagrammi a punti
- Diagrammi a linee
- Istogrammi
- Poligoni di frequenza relativa
- Curve di frequenza cumulata

Metodi Esplorativi

- Stem and leaf
- Box plot

Metodi per investigare Il livello di associazione

Tra variabili diverse

- Diagrammi di correlazione
- Q-Q plot

Di una variabile Con sé stessa

Definizione di frequenza empirica

Frequenza assoluta: La frequenza empirica assoluta di una certa caratteristica è data dal numero di volte che essa si presenta all'interno di un dato campione (operativamente: serie di dati)

$$x_i \rightarrow n_i$$

Frequenza relativa: La frequenza empirica relativa è definita come il Rapporto fra la frequenza assoluta di una certa Caratteristica e la numerosità del campione n (numero di dati totale)

$$f_i \rightarrow \frac{n_i}{n}$$

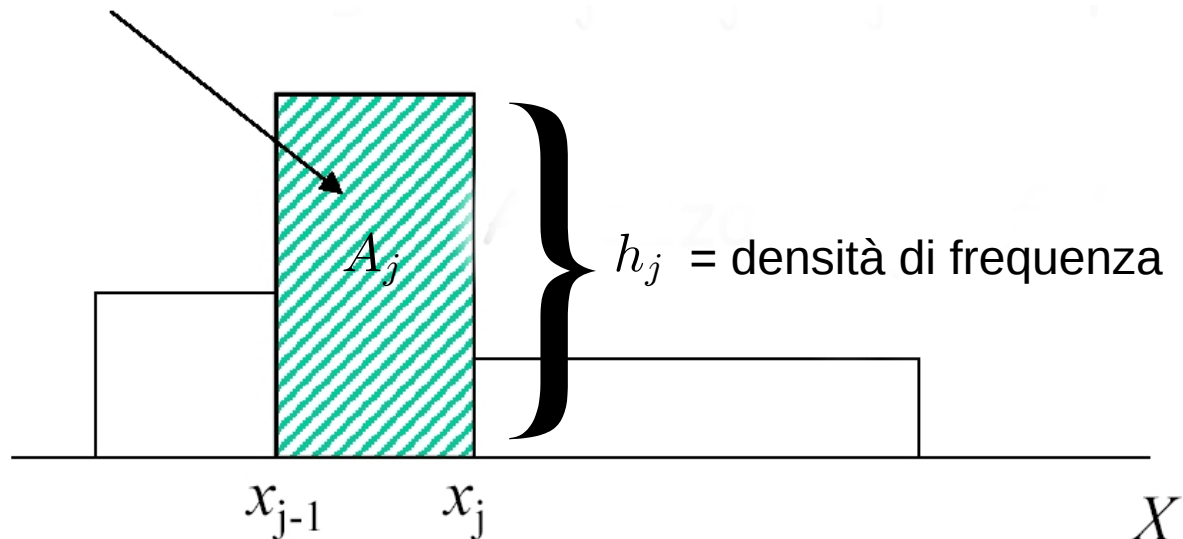
Per cui si ha ovviamente che:

$$\sum_{i=1}^n n_i = n \quad \text{e} \quad \sum_{i=1}^n f_i = 1$$

Metodi puramente descrittivi

Istogrammi

$$A_j = (x_j - x_{j-1}) \cdot h_j = \Delta x \cdot h_j = \frac{n_j}{n} = f_j$$



Alcune espressioni per la scelta del numero delle Classi n_c

$$n_c = \sqrt{n} \quad (1)$$

$$n_c = 1 + 3.3 \log_{10} n \quad (2)$$

Range dei dati

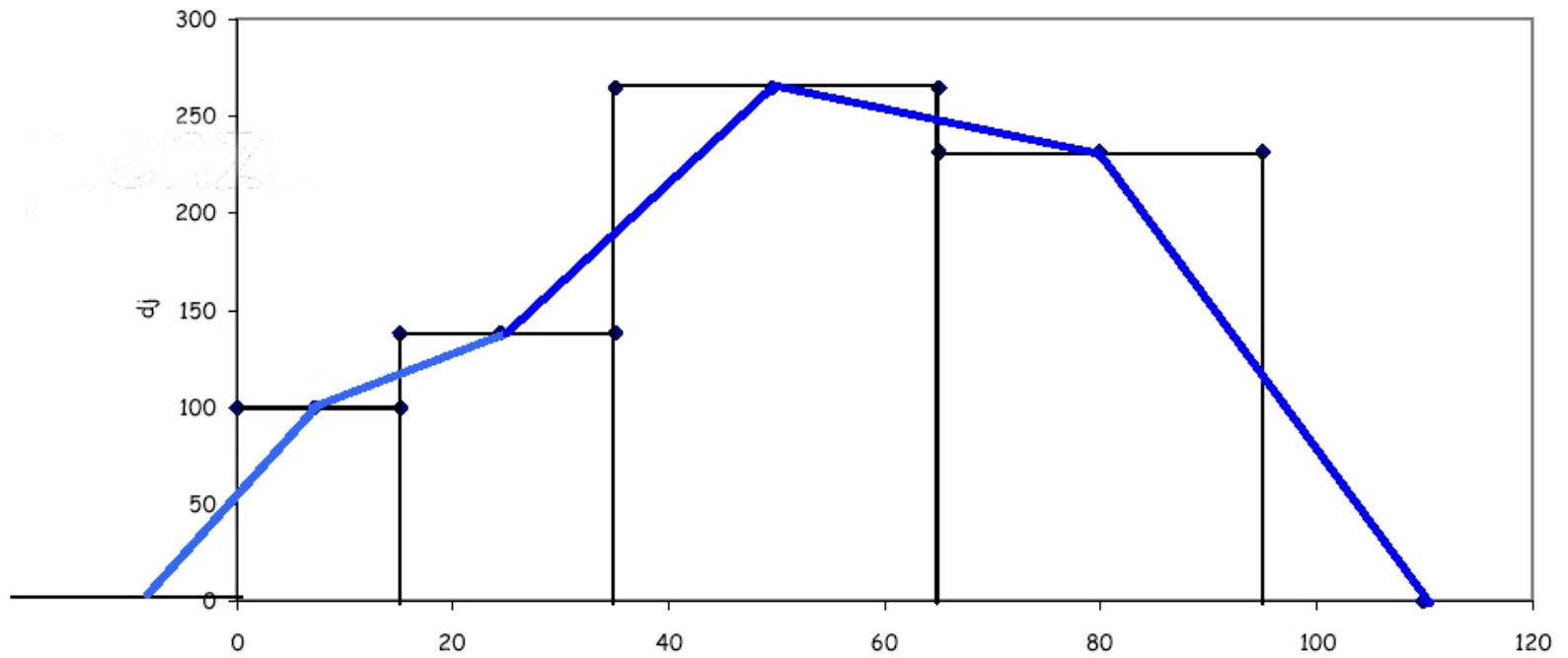
$$n_c = \frac{r n^{1/3}}{2iqr} \quad (3)$$

Distanza inter-quartile

Con:

$$iqr = Q_3 - Q_1$$

Poligoni di frequenza



Funzione di ripartizione empirica

Si definisce **funzione di distribuzione cumulata empirica** o **funzione di ripartizione empirica** di una variabile X , e si indica con F_X quella applicazione:

$$F_X : R \rightarrow [0, 1]$$

tale che

$$F_X(x) = P[X \leq x] = P[u : X(u) \leq x] \quad \forall x \in R$$

Alcune proprietà della funzione di ripartizione empirica

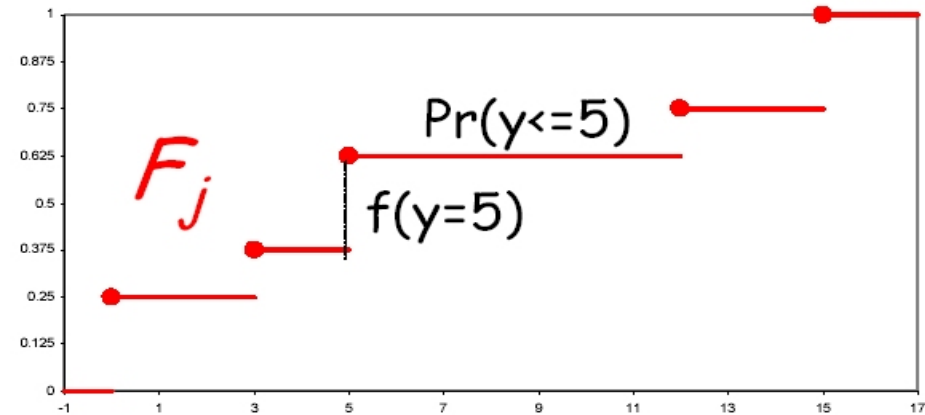
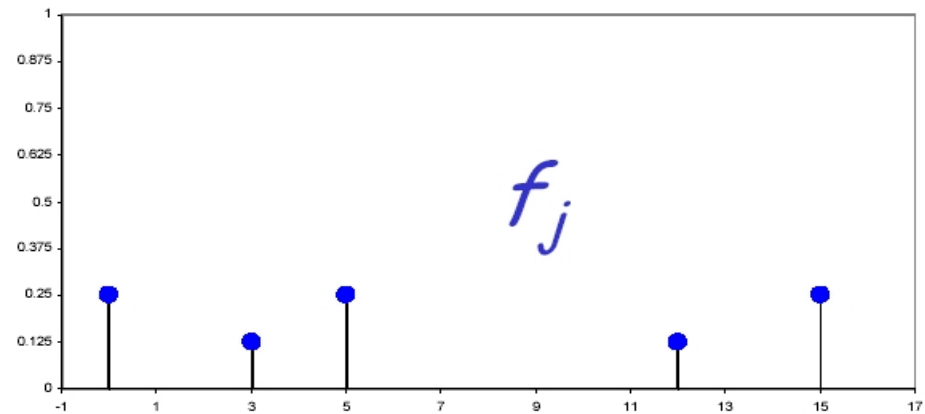
$$\lim_{x \rightarrow -\infty} F_X(x) = 0 \quad \text{e} \quad \lim_{x \rightarrow +\infty} F_X(x) = 1$$

F_X è monotona non decrescente cioè per $a < b$ $F_X(a) \leq F_X(b)$

F_X è continua da destra cioè: $\lim_{h \rightarrow 0^+} F_X(x + h) = F_X(x)$ con $h > 0$

Un esempio:
Funzione di ripartizione
Empirica per una variabile
discreta

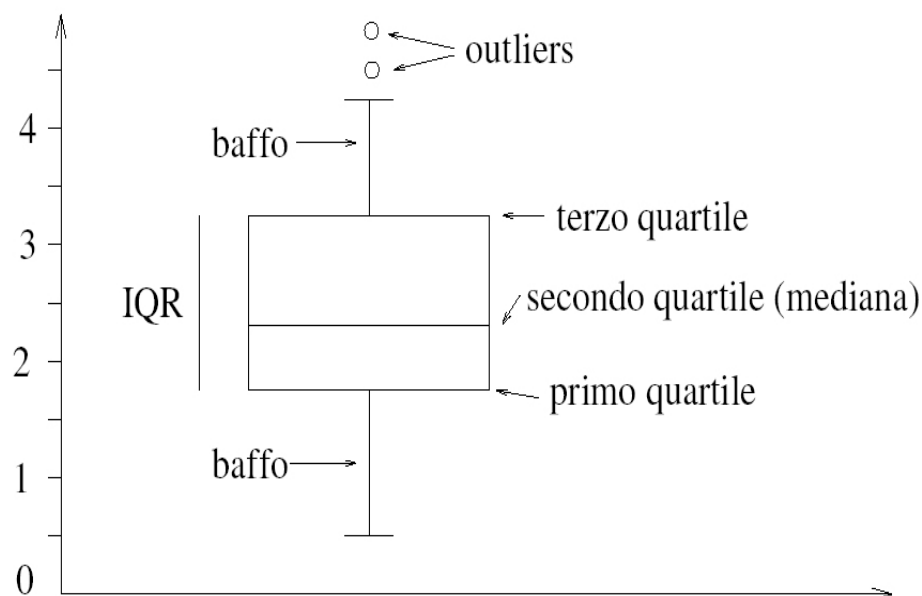
$$F_X(x) = P[X \leq x] = P[u : X(u) \leq x]$$



Definizione di quantili

Se si denota con q un dato livello di frequenza relativa, Il q -esimo quantile È il più piccolo numero ξ che soddisfa la disequaglianza:

$$F_X(\xi) \geq q$$



Box-plot

Riassunti numerici dei dati

Misure di tendenza centrale

- media
- moda
- medianac

Misure di dispersione

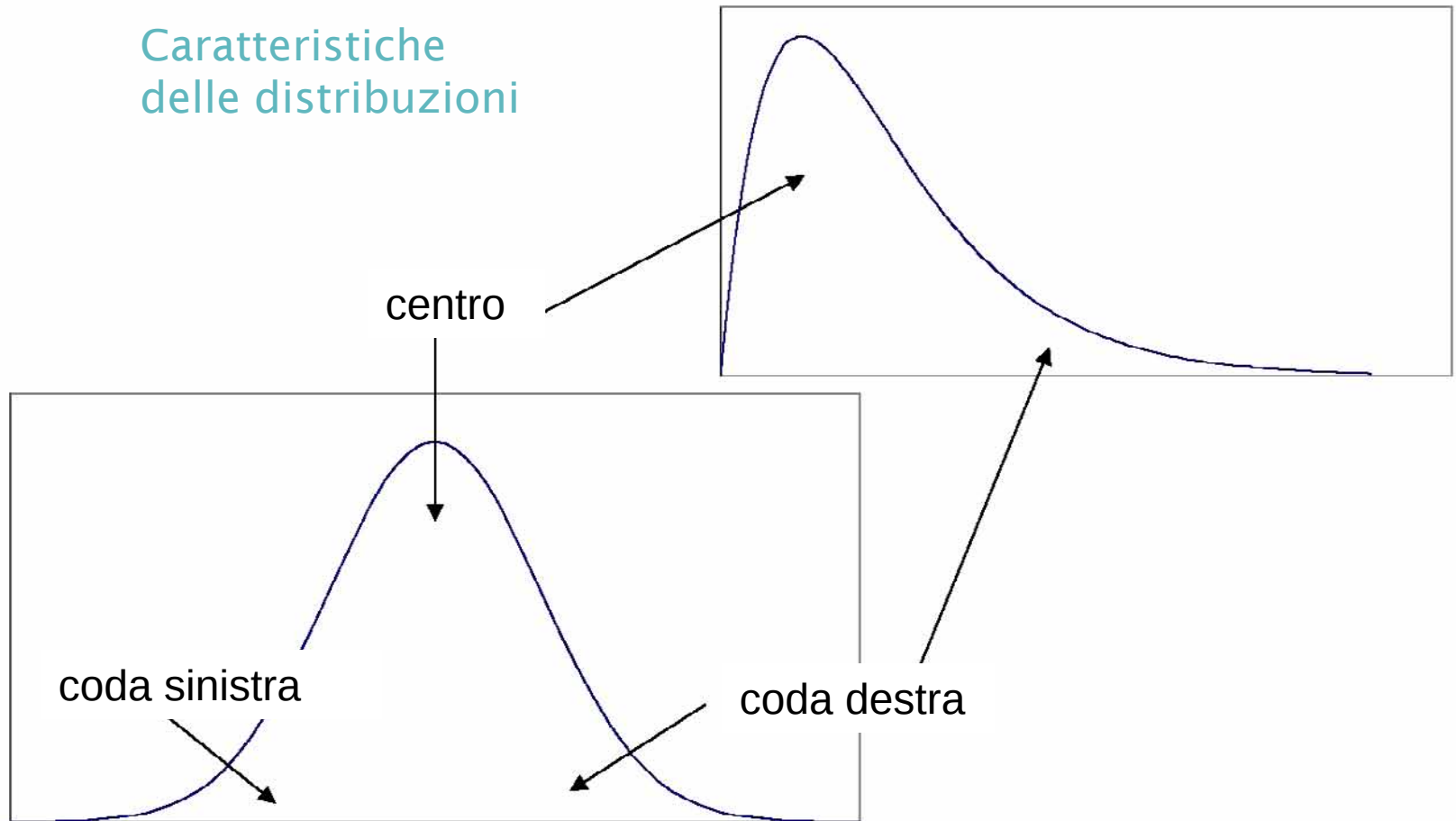
+

Coefficiente di curtosi:
altezza relativa del picco
richiede un campione ampio
per distribuzioni simmetriche

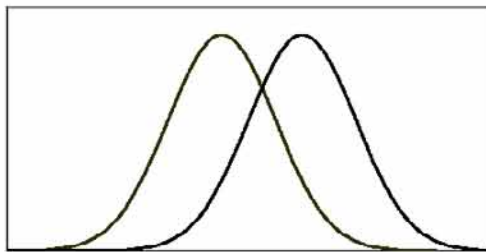
Misure di asimmetria

Coefficiente
di asimmetria

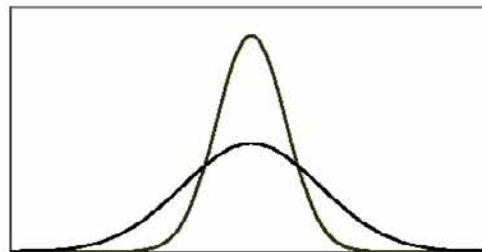
Caratteristiche delle distribuzioni



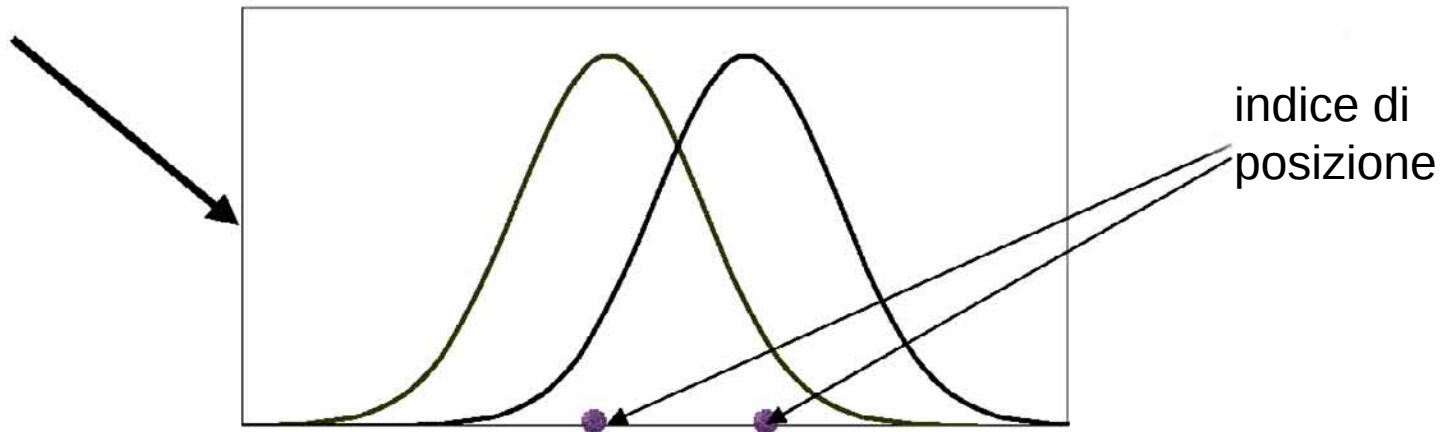
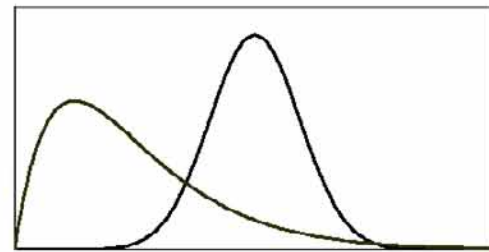
posizione



variabilità



simmetria



Misure di tendenza centrale

Media empirica

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

Media Spuntata (trimmed mean)

Calcolata considerando solo il 90% centrale dei Dati (cioè compresi tra il 5% ed il 95% dei dati ordinati)

moda

Valore/i con frequenza massima

mediana

E' il valore dell'osservazione per cui nel campione ci sono il 50% delle osservazioni minori o uguali a questa

La media empirica, essendo il baricentro dei dati, risente molto della posizione Dei valori estremi

La mediana invece, non è assolutamente influenzata dagli estremi

Misure di dispersione

Varianza empirica:

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

Uno stimatore più robusto della varianza è:

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

Scarto:
$$\sigma = \left[\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 \right]^{\frac{1}{2}}$$

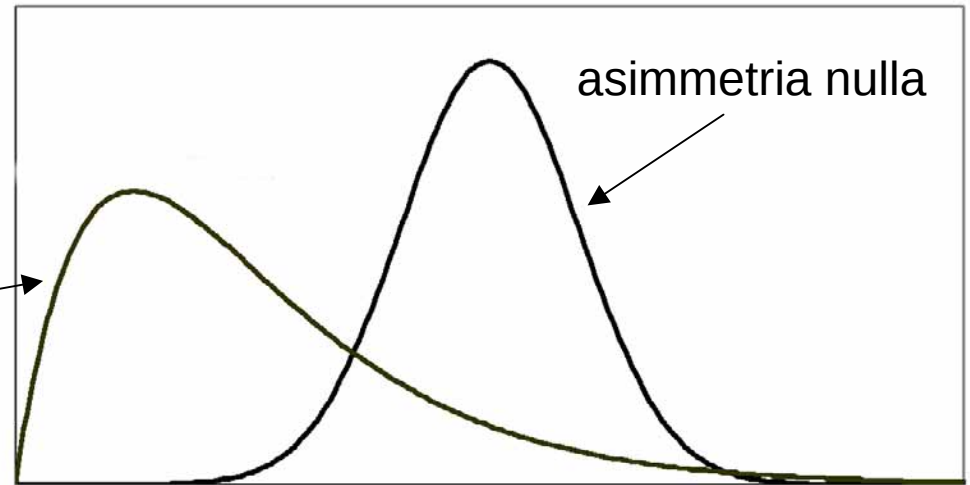
Scarto spuntato...

Range: (valore massimo-valore minimo)

Coefficiente di variazione: $cv = \frac{s}{\bar{x}}$ con $\bar{x} \neq 0$

Misure di Asimmetria

asimmetria positiva



Coefficiente di asimmetria
o skewness

$$g_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})^3}{n s^3}$$

misura l'asimmetria
rispetto alla media

Coefficiente di curtosi
o peakedness

$$g_2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^4}{n s^4}$$

misura il "peso
delle code"