PROBABILE... ?

Se noi non fossimo ignoranti non ci sarebbe probabilità,
Ci potrebbero essere solo certezze. Ma la nostra ignoranza
non può essere assoluta, altrimenti non ci sarebbe più
Probabilità. Così i problemi di probabilità possono
essere classificati a seconda della maggiore o minore
Profondità della nostra ignoranza.

(H. Poincarè)

PREDICIBILE... ?

-Le previsioni del tempo – disse Mr Oliver, voltando le pagine finché non le ebbe trovate, - dicono: venti variabili; temperatura moderata; piogge occasionali... -. C'era una certa irresponsabilità, un'assenza di ordine e di simmetria nelle nuvole, nel loro diradarsi ed addensarsi. Obbedivano ad una legge propria, oppure non ubbidivano ad alcuna legge?

(Virginia Woolf – Tra un atto e l'altro)

Parlando di **Idrologia 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)

Un po' di bibliografia ...

- *Probability, Random Variables and Stochastic Processess,* A. Papoulis, S.U. Pillai, McGraw-Hill, 2002
- *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. Cornell, McGraw-Hill,1970

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

http://www.diam.unige.it/costid/acquedottiefognature.htm

Statistica: Scienza delle decisioni in condizioni di incertezza…

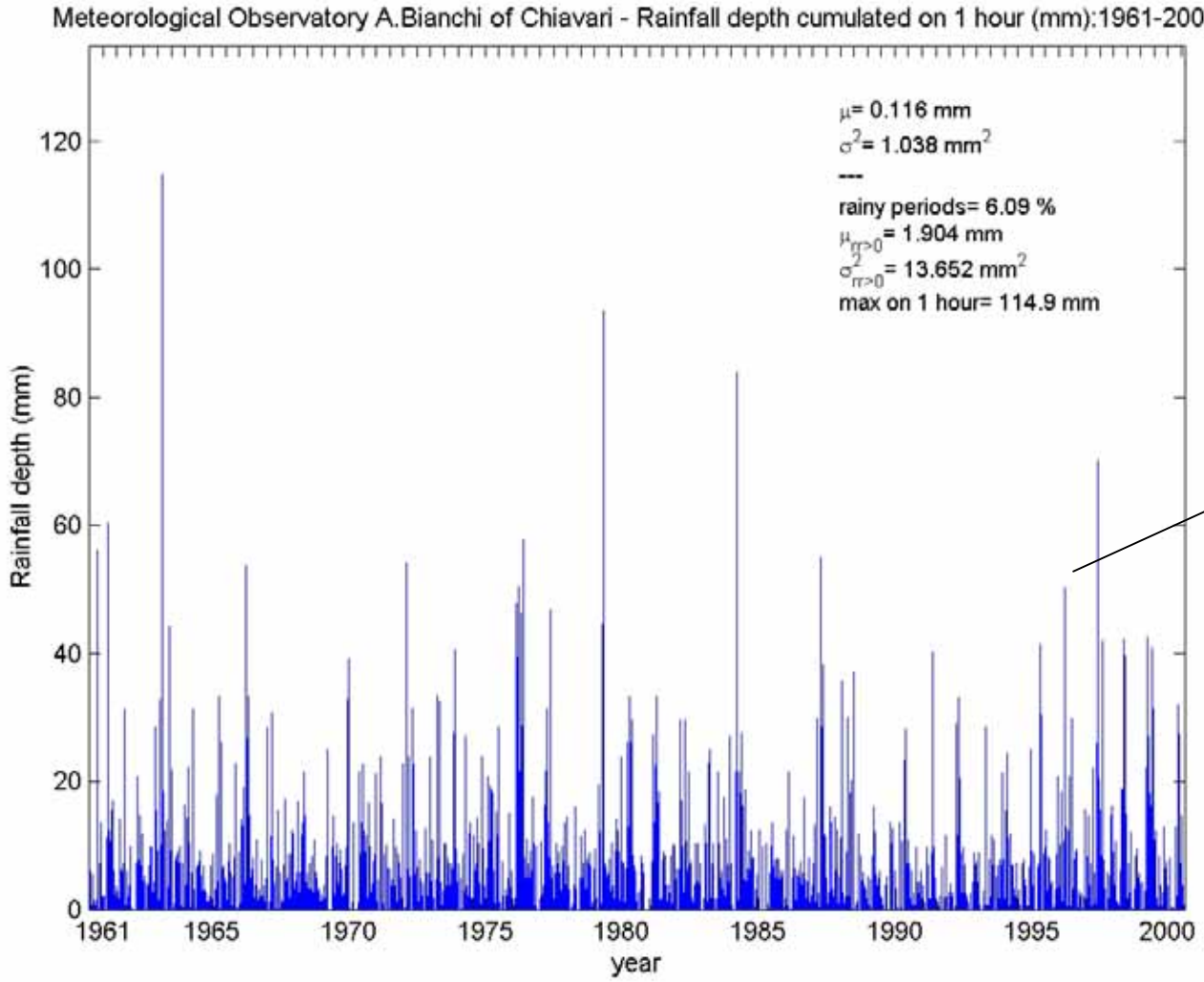
A che tipo di fenomeni si applica?

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



Elevata Variabilità

↓

Intermittenza

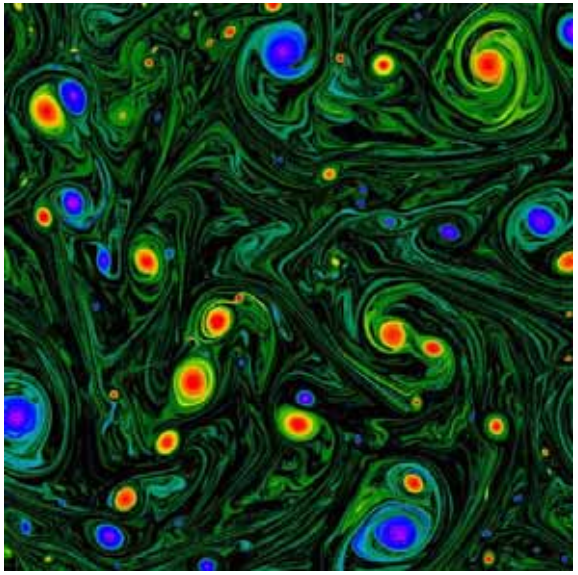
↓

Impredicibilità

Processi Deterministici

Processi Casuali

Chaos Deterministico?



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.

$U = \{u_i\}_{i=1}^{N}$ con N = 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

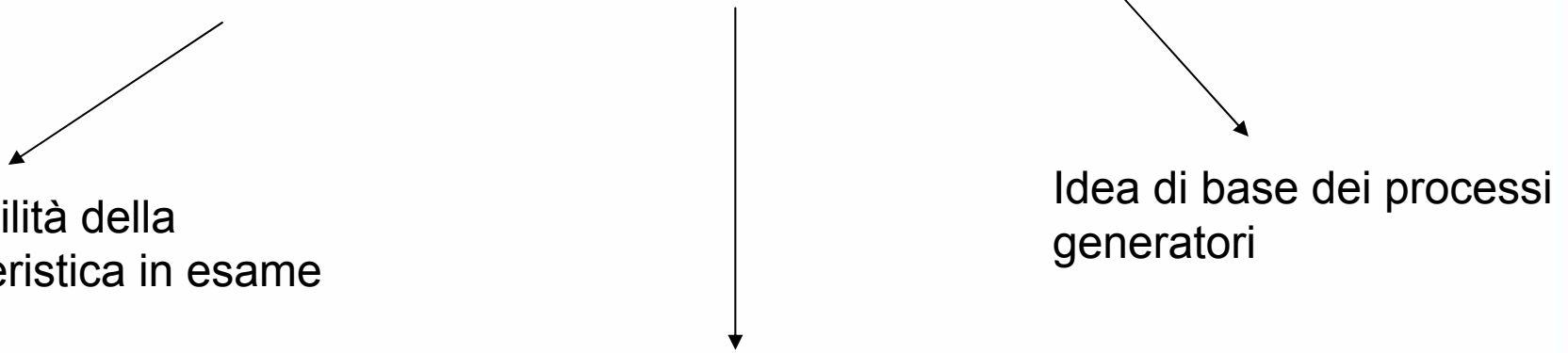
L'insieme delle modalità è finito o numerabile

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



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)

$$x_i \longrightarrow n_i$$

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)

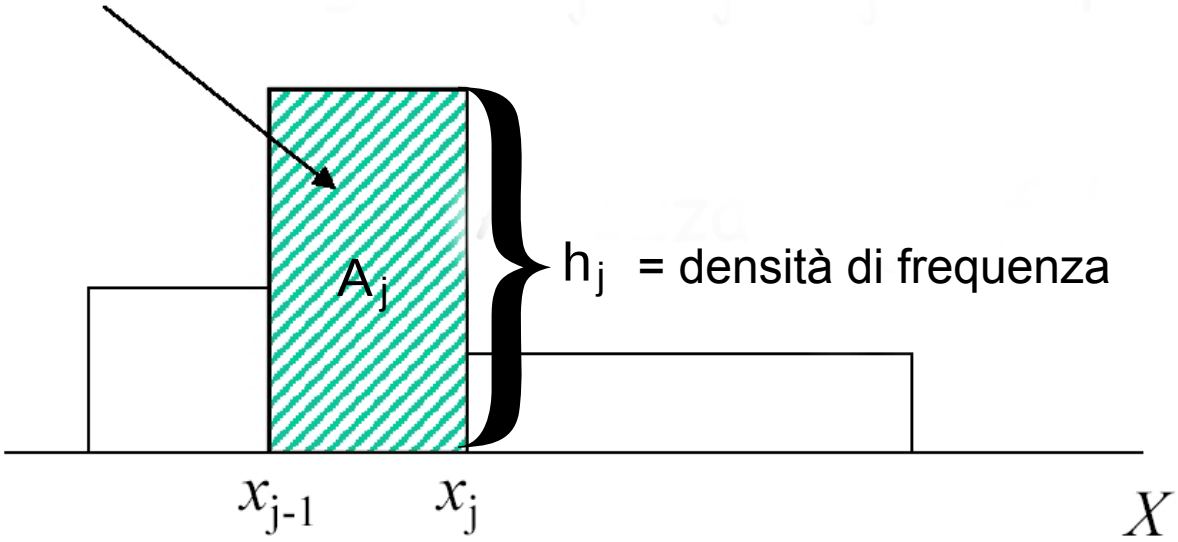
$$f_i \longrightarrow \frac{n_i}{n}$$

Per cui si ha ovviamente che: $\sum_{i=1}^{p} n_i = n$ e $\sum_{i=1}^{p} f_i = 1$

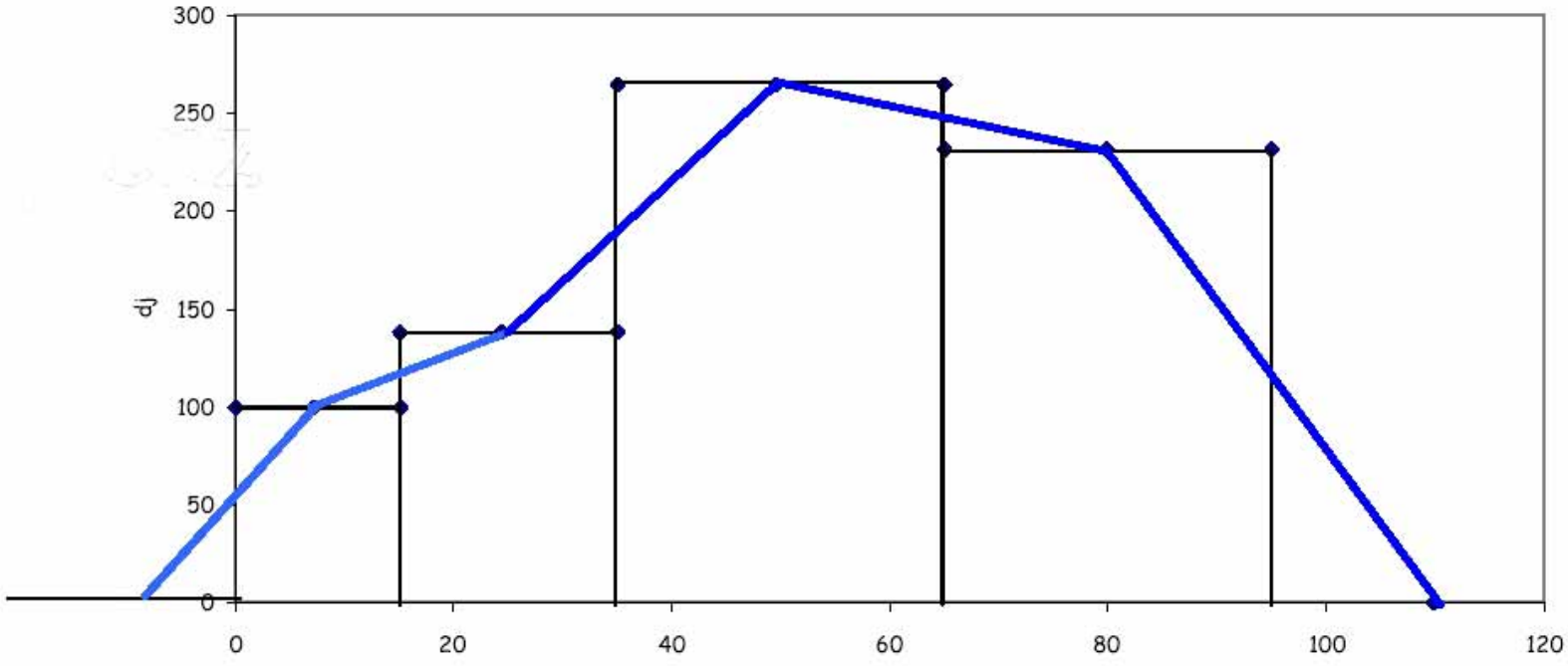
Metodi puramente descrittivi

Istogrammi

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



Poligoni di frequenza



Al limite (per la larghezza degli intervalli dell'istogramma che tende a zero)
-> FUNZIONE DENSITA' DI PROBABILITA'

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

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

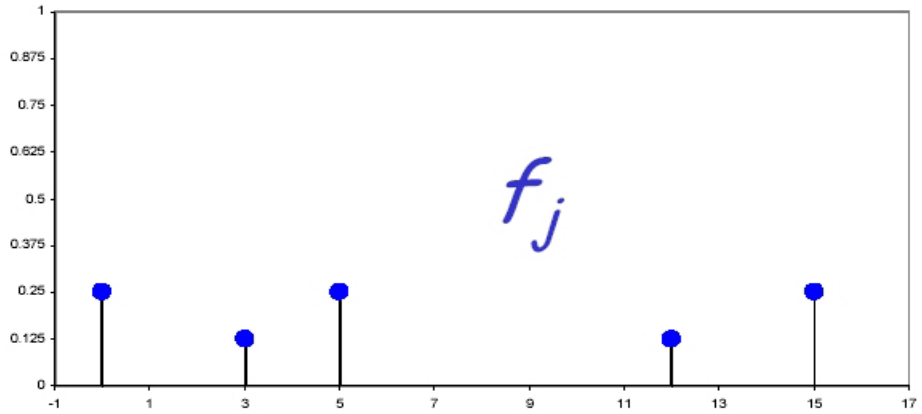
Alcune proprietà della funzione di ripartizione empirica

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

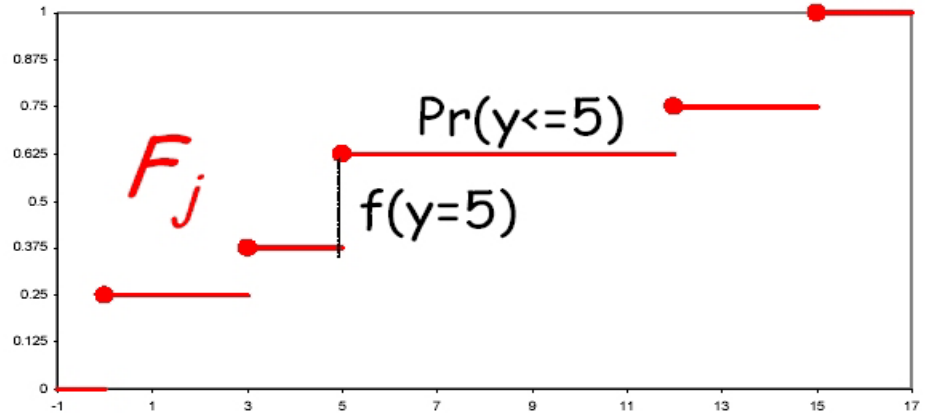
F_X è <u>monotona non decrescente</u> cioè per $a < b \quad F_X(a) \leq F_X(b)$

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

Un esempio:
Funzione di ripartizione
Empirica per una variabile discreta



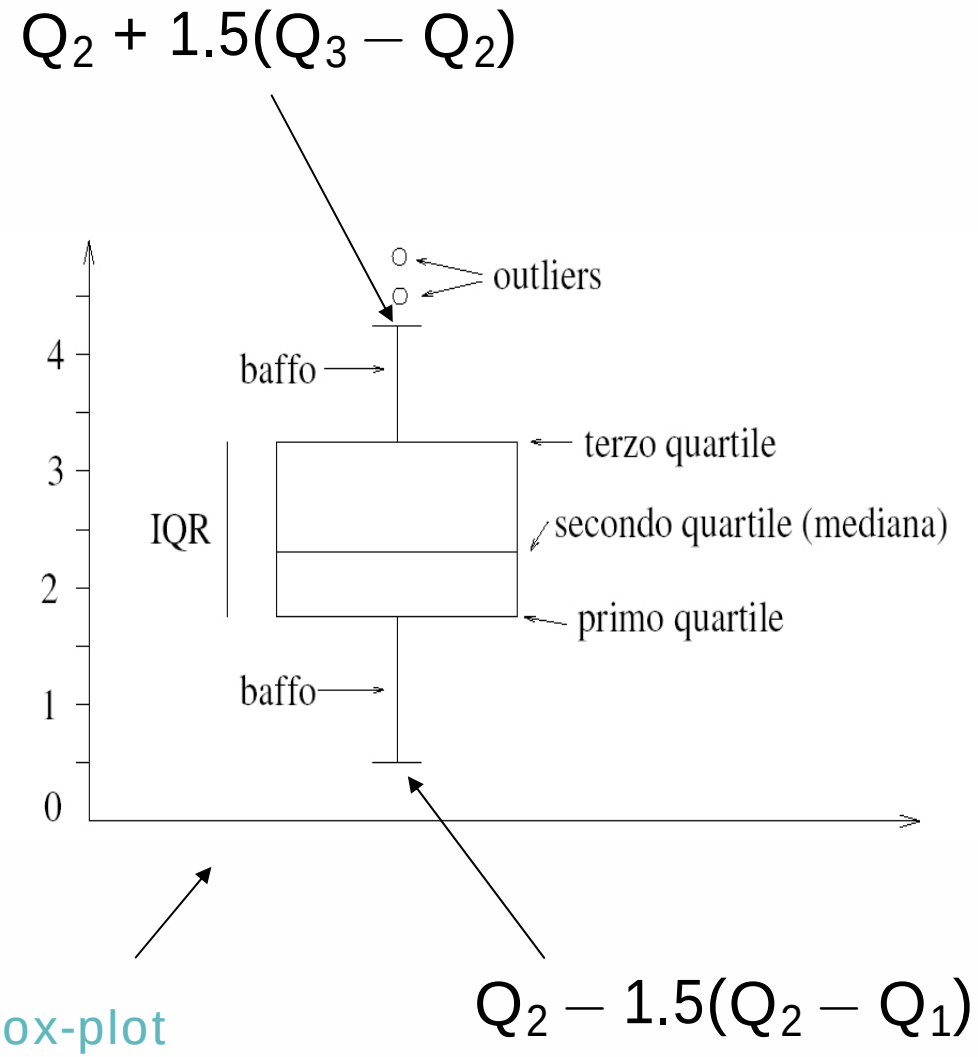
$$F_X(x) = P[X \leq x] = P[u : X(u) \leq x]$$



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 diseguaglianza:

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



Box-plot

Un po' di nozionismo (in formule)...

Numero ideale di classi di un istogramma (regola "di massima"):

$$n_c = \sqrt{n}$$

Sturges (1926):

$$n_c = 1 + 3.3 \log_{10} n$$

Freedman e Diaconis (1981):

$$n_c = \frac{r n^{\frac{1}{3}}}{2 i r q}$$

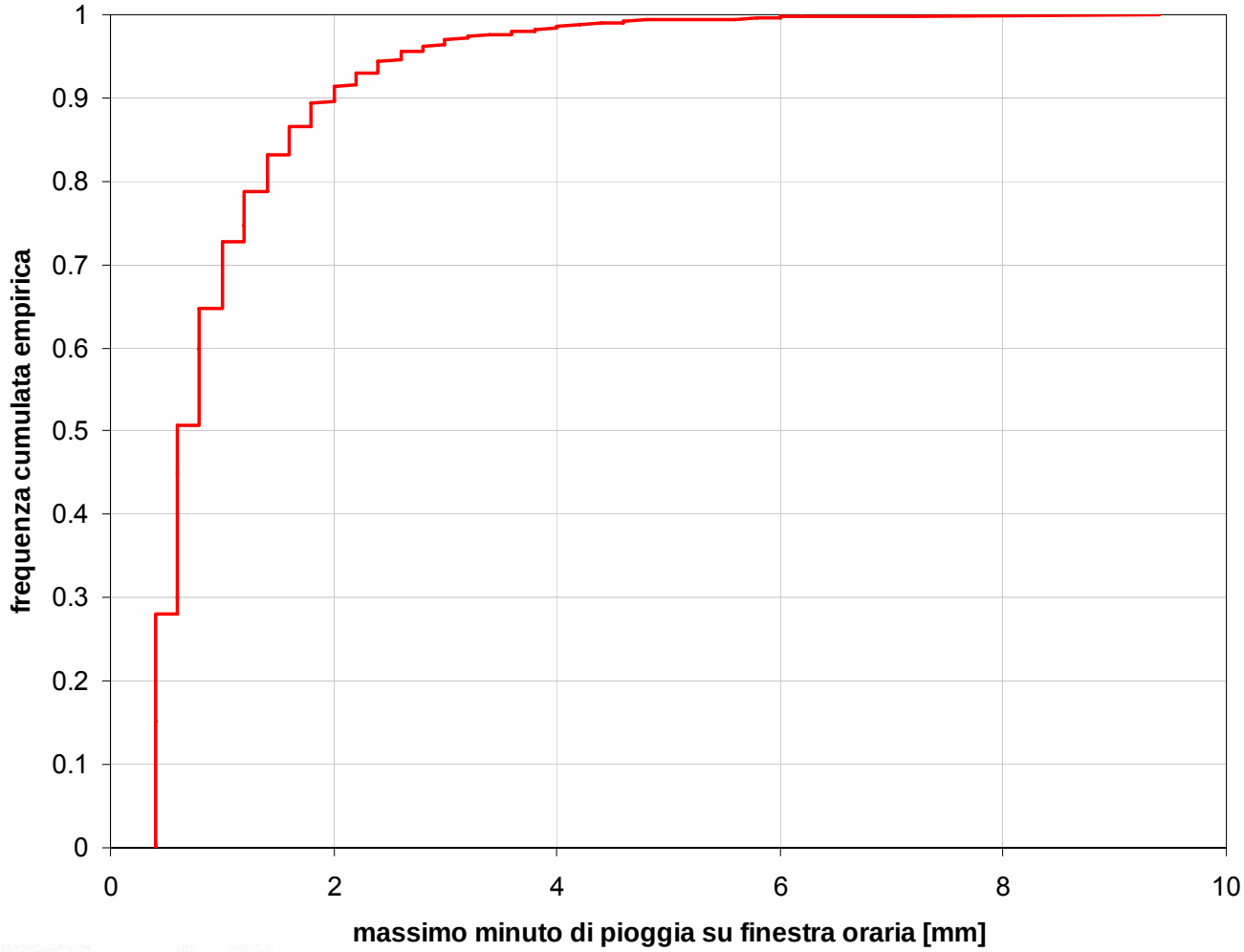
Distanza inter-quartile

$$irq = Q_3 - Q_1$$

range

Diagramma di frequenza cumulata (o frequenza cumulata empirica)

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



Plotting-position e verifica dell'adattamento di una legge di distribuzione di probabilità ad un campione di dati (1)

Dato quindi un campione di dati, sufficientemente numeroso, è possibile tracciare la curva di frequenza empirica effettuando le seguenti operazioni:

- disporre i dati in ordine crescente
- associare ad ogni valore il numero d'ordine ***i***
- stimare la frequenza empirica di non superamento F_i

ricordando inoltre che la frequenza empirica di non superamento dell'elemento x_i, è data dal numero di elementi del campione che hanno valore minore o al più uguale ad x_i, si può scrivere:

$$F_i = \frac{i}{N}$$

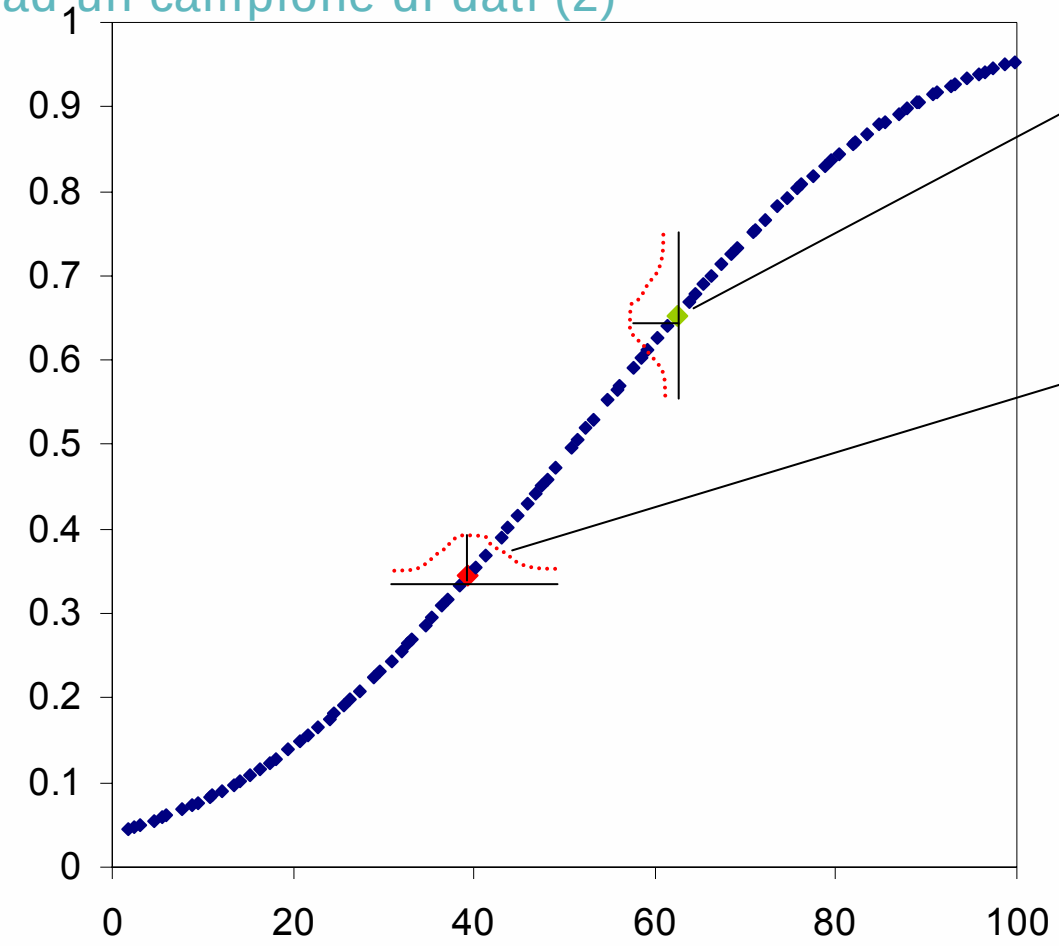
Frequenza dell'ultimo Valore pari a 1

Inacettabile statisticamente eccetto che per distribuzioni limitate superiormente

$$F_i = \frac{i}{N+1}$$

Plotting-position di Weibull

Plotting-position e verifica dell'adattamento di una legge di distribuzione di probabilità ad un campione di dati (2)



$$F_i = \frac{i}{N+1}$$

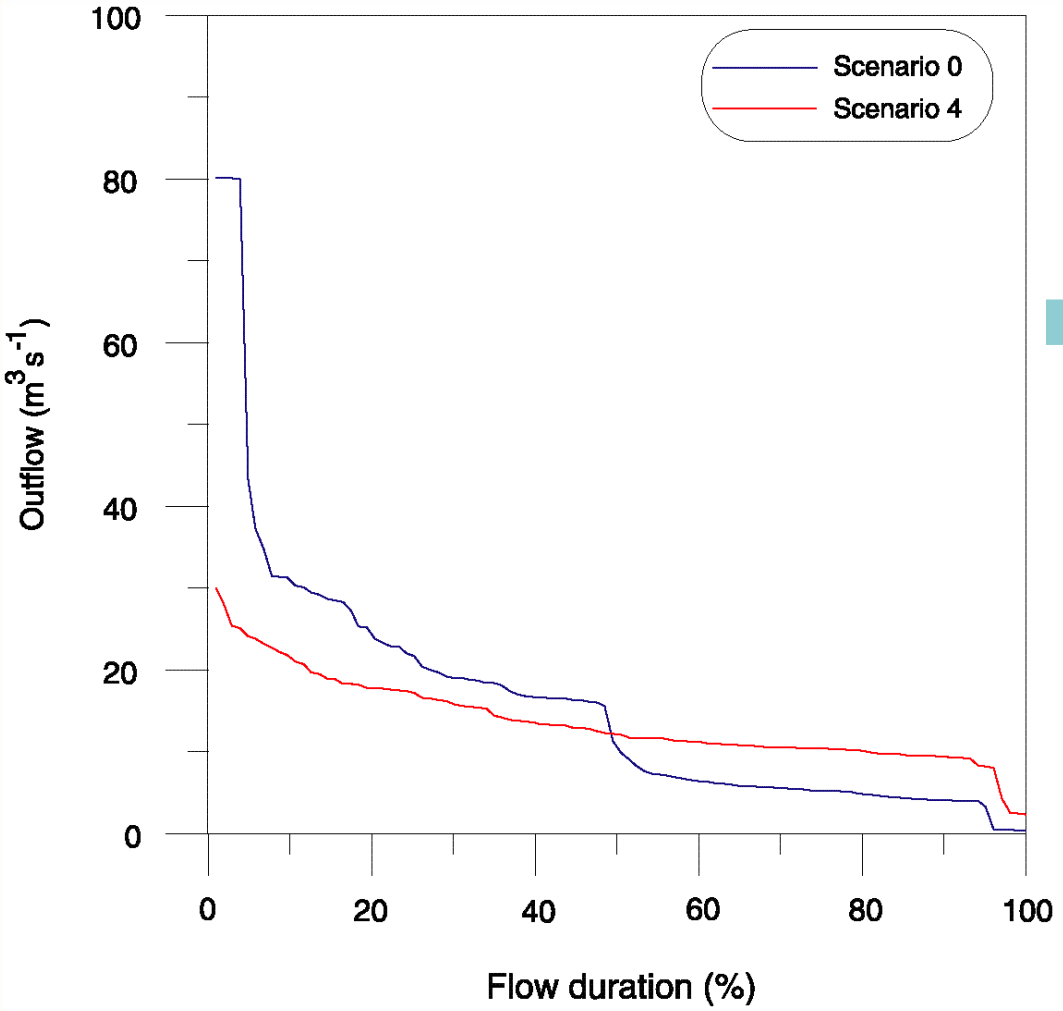
Plotting-position di Weibull

$$F_i = \frac{i-\ddot{e}}{N+1-2\ddot{e}}$$

Il coefficiente α assume i valori

- *0.4 per distribuzione debolmente simmetrica*
- *0.44 per distribuzione mediamente asimmetrica*
- *0.5 per distribuzione fortemente asimmetrica*

ottenendo rispettivamente le formule proposte da Cunnane, Gringorten, Hazen



Curve di durata

forniscono il numero medio di giorni all'anno (o la media percentuale di giorni all'anno) in cui una certa portata viene superata

$$D_Q(q) = 365 \cdot [1 - F_Q(q)]$$

Riassunti numerici dei dati

Misure di tendenza centrale

- media
- moda
- mediana

Misure di dispersione

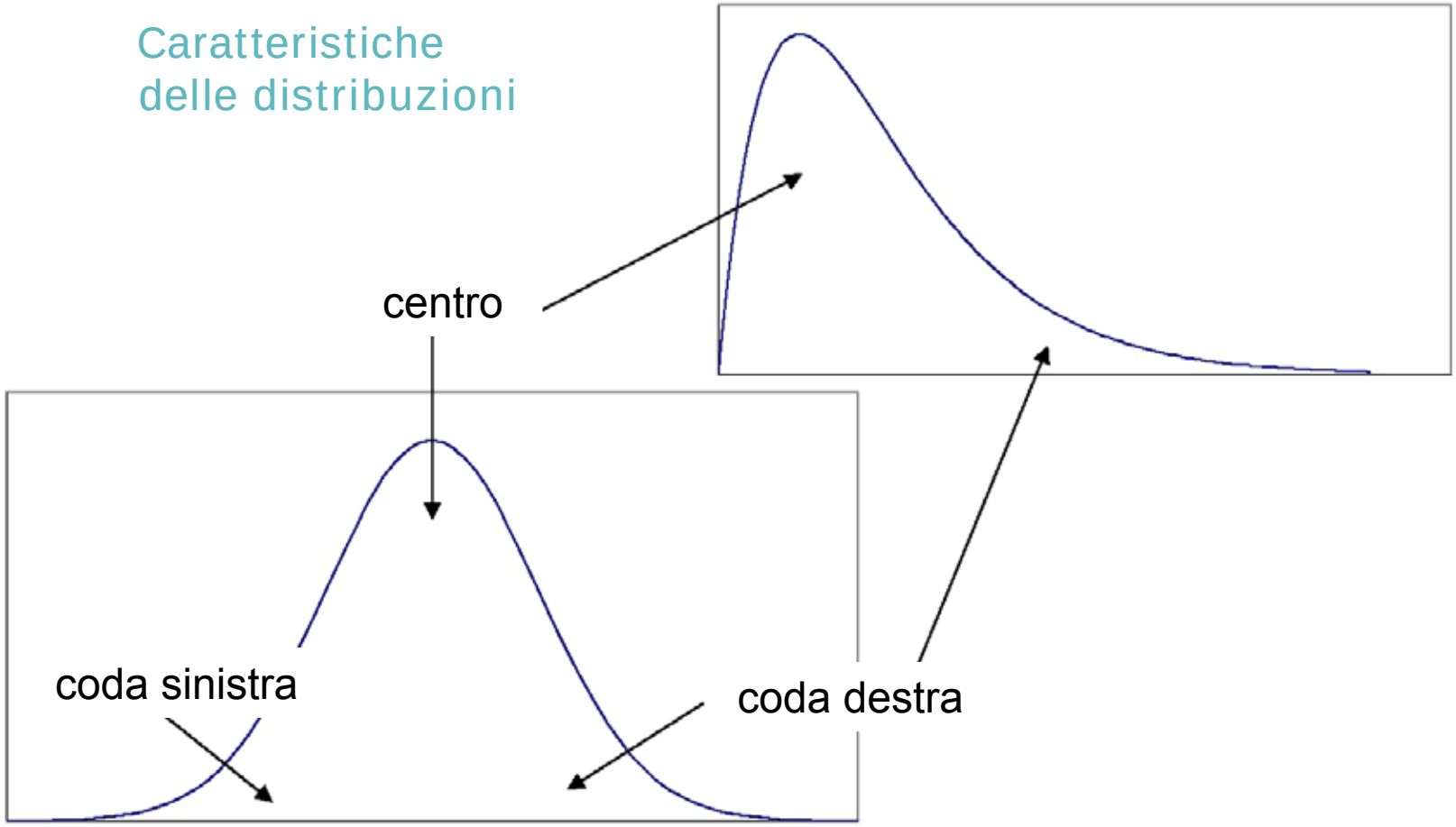
+

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

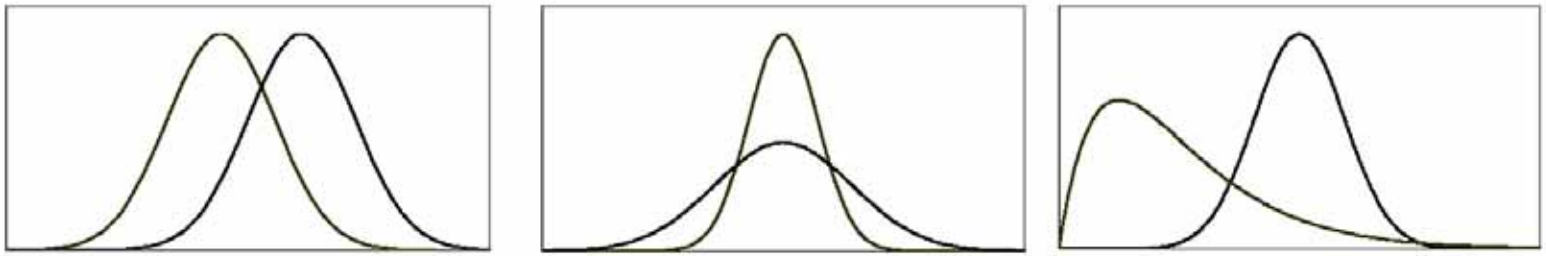
Misure di asimmetria

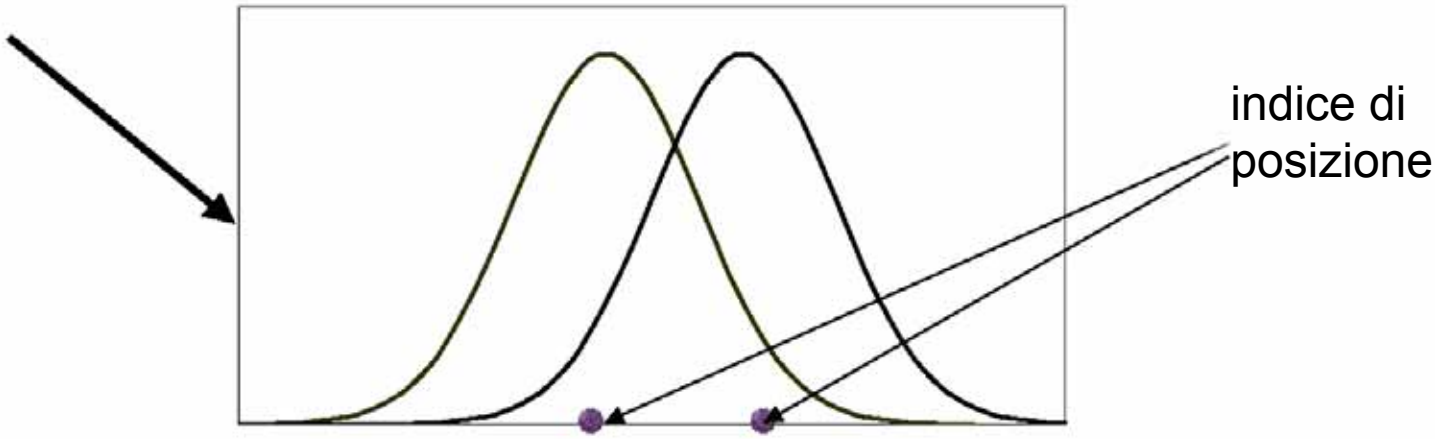
Coefficiente
di asimetria

Caratteristiche delle distribuzioni



posizione variabilità simmetria





Misure di tendenza centrale (1)

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 tendenza centrale (2)

Media armonica $\bar{x}_h = \frac{1}{(1/n)\cdot[(1/x_1)+(1/x_2)+\ldots+(1/x_n)]}$

Si applica quando ad essere mediato è il reciproco di una variabile

Media geometrica $\bar{x}_g = (x_1 x_2 \ldots x_n)^{\frac{1}{n}} = \exp\left[\frac{1}{n}\sum_{i=1}^{n}\log x_i\right] = \prod_{i=1}^{n} x_i^{\frac{1}{n}}$

Variabili che rappresentano tassi di incremento o decremento…

Misure di dispersione

Varianza empirica:

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

$$\hat{\sigma}^2 = E\left[X^2\right] - (E[X])^2$$

Uno stimatore più robusto della varianza è:

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

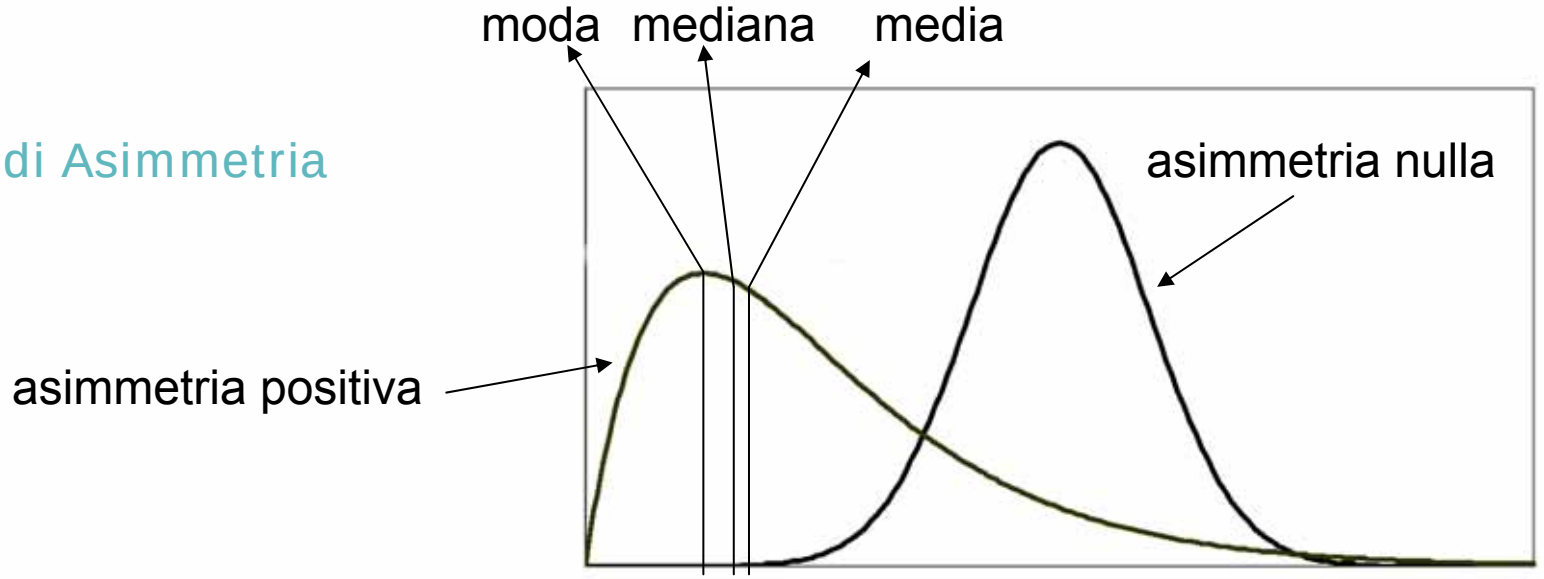
Scarto: $$\hat{\sigma} = \left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^{\frac{1}{2}}$$

Scarto spuntato…

Range: (valore massimo-valore minimo)

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

Misure di Asimmetria



Coefficiente di asimmetria o skewness

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

misura l'asimmetria rispetto alla media

Coefficiente di curtosi o peakedness

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

misura il "peso delle code"