# METODI STATISTICI PER L'ANALISI DI DATI FINANZIARI MEDIANTE IL SOFTWARE R


**Paolo Giordani**

SAPIENZA
UNIVERSITÀ DI ROMA

paolo.giordani@uniroma1.it

**Centro Interdipartimentale**
**di Ricerca DigiLab**



**25 marzo 2019**

**Seminario organizzato dal Comitato Scientifico dell'Ordine degli Attuari**

## Problema

Il rendimento (futuro) $R$ di un investimento è una quantità affetta da incertezza

Si assume che questa incertezza sia di natura probabilistica

Indicando con $R_i$ il rendimento al tempo $i$ si ha che

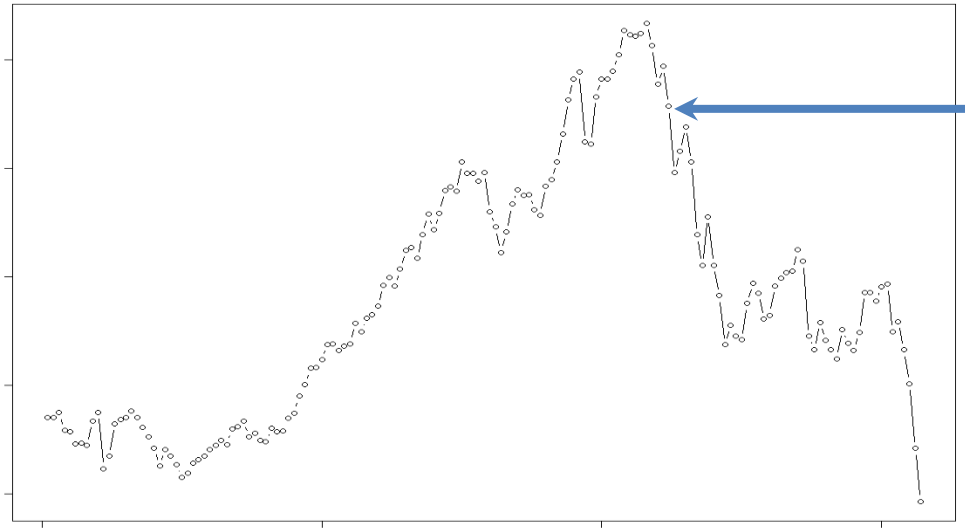
$R_i$ è una variabile aleatoria

Il problema da affrontare è l'individuazione del meccanismo aleatorio che sottende $R_i$

Al fine di rendere il problema più generale (qualsiasi grandezza di carattere finanziario), si utilizzerà nel seguito una differente notazione

$$R_i \longrightarrow Y_i$$

## Problema (in pratica)



valore osservato $y_i$
realizzazione di v.a. $Y_i$

$Y_i \sim$ ???
- che distribuzione?
- quali parametri?

Campione osservato di dimensione $n$ $\quad \underline{y} = y_1, \ldots, y_i, \ldots, y_n$

Realizzazione di $n$ v.a. $\quad \underline{Y} = Y_1, \ldots, Y_i, \ldots, Y_n$

Le $n$ v.a. $Y_1, \ldots, Y_i, \ldots, Y_n$ sono assunte indipendenti e identicamente distribuite (i.i.d.)

**Problema (da un punto di vista statistico)**

$$Y_i \ , i = 1, \ldots, n \sim \ F \text{ (con } F \text{ ignota)}$$

$F$ è la distribuzione marginale (distribuzione di $Y_i$ al netto di $Y_{i'}$ con $i' \neq i$)

Modello Statistico

$$\{\underline{y} \in \Upsilon, f_{\underline{Y}}(\underline{y}) = \prod_{i=1}^{n} f_p(y_i), p \in \mathrm{P}\}$$

dove $f_p$ è la f. di densità di $F$ associata alla legge di probabilità $p$ incognita

Modello Statistico Parametrico

$$\{\underline{y} \in \Upsilon, f_{\underline{Y}}(\underline{y}) = \prod_{i=1}^{n} f_\theta(y_i), \theta \in \Theta\}$$

dove $f_\theta$ è la funzione di densità di $F$ associata ad un parametro $\theta$ incognito

---

**"All models are wrong, but some are useful"**
**George E.P. Box**

## Assunzione di indipendenza

Test $H_0$: Indipendenza
$H_1$: Dipendenza

Ljung - Box

$$Q_{LB} = n\,(n+2)\sum_{k=1}^{h}\left[r_k^2/(n-k)\right]$$

Box - Pierce

$$Q_{BP} = n\sum_{k=1}^{h} r_k^2$$

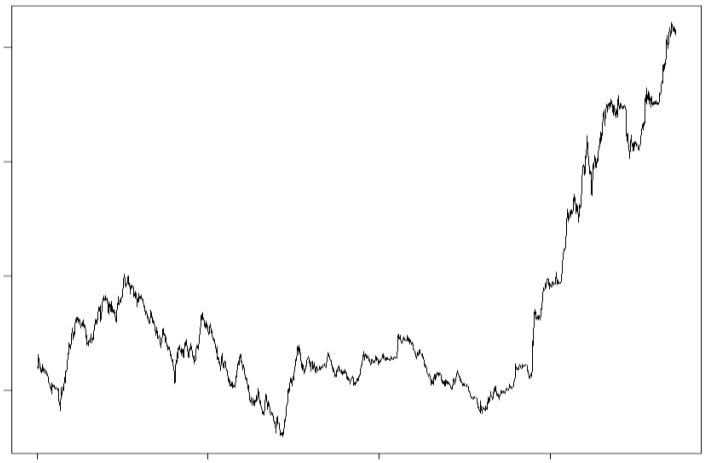
dove: $r_k$ indica l'autocorrelazione campionaria di ritardo $k$
$h$ indica il massimo ritardo considerato

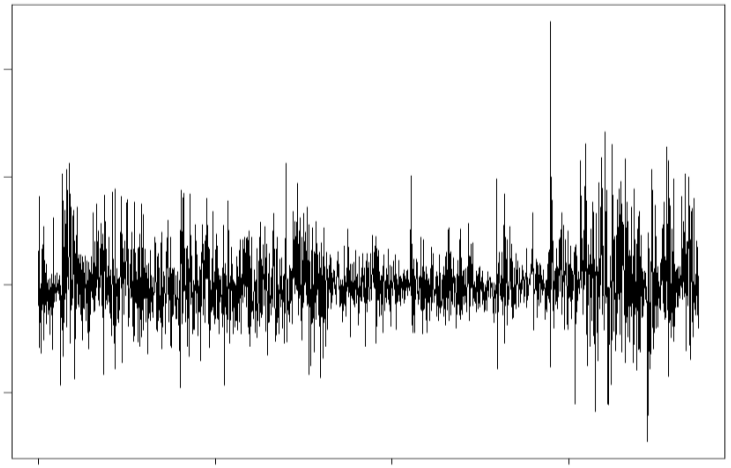
`stats::Box.test`

# Assunzione di identica distribuzione

"You can observe a lot by watching"
Lawrence Peter "Yogi" Berra

`graphics::plot`



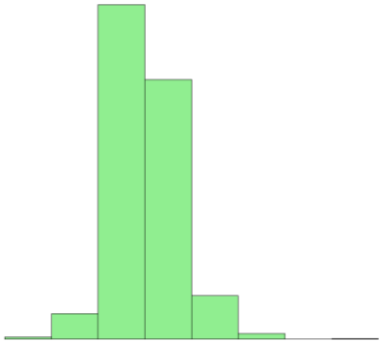


## Ricerca del modello statistico (graficamente)

Istogramma (classi della stessa ampiezza)

$$f_{\text{HIST}}(y) = 1/n \sum_{i=1}^{n} \delta_{\text{HIST}}(y, y_i)$$

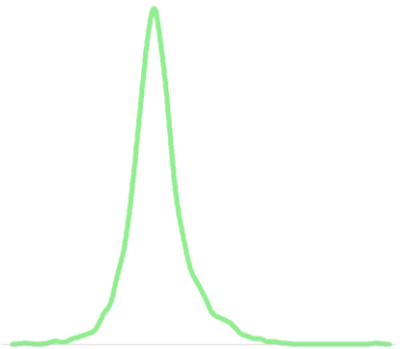
$$\text{con } \delta_{\text{HIST}}(y, y_i) = \begin{cases} 1 & \text{se } y \text{ e } y_i \text{ sono nella stessa classe} \\ 0 & \text{altrimenti} \end{cases}$$



Stima Kernel ($b$ ampiezza di banda – bandwidth, $K$ funzione Kernel)

$$f_{\text{KERNEL}}(y) = 1/n \sum_{i=1}^{n} \delta_{\text{KERNEL}}(y, y_i)$$

$$\text{con } \delta_{\text{KERNEL}}(y, y_i) = 1/b\, K[(y - y_i)/b] \in [0, 1]$$



`graphics::hist`

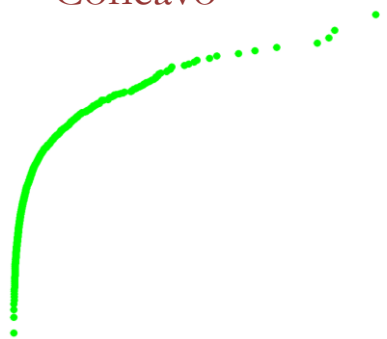
`stats::density`

## Selezione del modello statistico (graficamente)

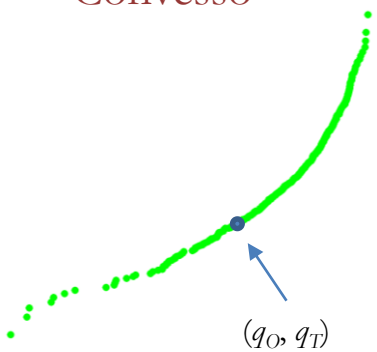
Q-Q qplot: quantili osservati (asse $x$) vs quantili teorici (asse $y$) – linearità?

$q_O < q_T$
Asimmetria a destra (della distribuzione campionaria rispetto a quella teorica)

Concavo



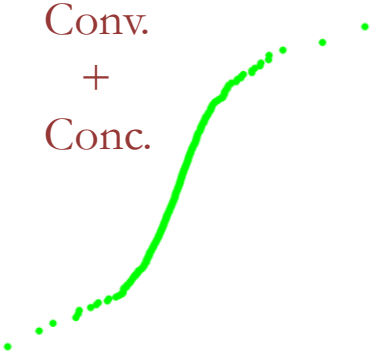
Convesso



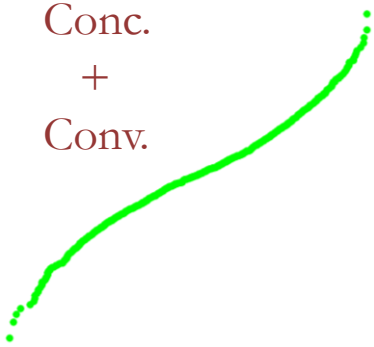

$q_O > q_T$
Asimmetria a sinistra (es.: $q_O = 40°$, $q_T = 20°$, il 20° quantile teorico corrisponde al 40° quantile osservato; massa sulla coda sinistra)

Mix dei precedenti:
Simmetria,
ma code pesanti

Conv. + Conc.



Conc. + Conv.



Mix dei precedenti:
Simmetria,
ma code leggere

`stats::qqnorm`

`stats::qqplot`

## Normalità?

Test

Shapiro - Wilk (correlazione tra quantili teorici e osservati)

Kolmogorov - Smirnov (funzione di ripartizione teorica e osservata)

Cramer - von Mises (funzione di ripartizione teorica e osservata)

Jarque - Bera (curtosi e simmetria)

```
stats::shapiro.test
stats::ks.test
nortest::cvm.test
tseries::jarque.bera.test
```

## Stima

Metodo della Massima Verosimiglianza

$$\theta_{MV} = \underset{\theta}{\text{argmax}}\ L(\theta) = f_{\underline{Y}}(\underline{y}) = \prod_{i=1}^{n} f_{\theta}(y_i)$$

(Alcune) proprietà asintotiche dello stimatore di massima verosimiglianza


Consistenza


Non distorsione


Efficienza


Normalità

```
MASS::fitdistr
nor1mix::norMixEM
...
stats::optim
stats::nlminb
```

## Selezione del modello statistico (indicatori)

Parsimonia: modello più semplice in grado di spiegare "bene" il fenomeno

- un modello troppo semplice (pochi parametri) non è in grado di cogliere alcune caratteristiche rilevanti del fenomeno.

- un modello troppo complesso (tanti parametri) aumenta l'incertezza (errore standard dei parametri da stimare) ed è più difficile da interpretare

Indicando con $p$ il numero dei parametri

$$\text{AIC} = -2 \ln[L(\theta)] + 2p$$

$$\text{BIC} = -2 \ln[L(\theta)] + p \ln(n)$$

```
stats::AIC
stats::BIC
```