

Università di Siena

Corso di STATISTICA

Parte prima: Probabilità e variabili aleatorie

Andrea Garulli, Antonello Giannitrapani, Simone Paoletti

Master E^2C

Centro per lo Studio dei Sistemi Complessi

Università di Siena

email: statistica@csc.unisi.it

Lezione # 1

- ✓ Teoria della probabilità
- ✓ Variabili aleatorie
- ✓ Distribuzioni di probabilità
- ✓ Distribuzioni bivariate e multivariate
- ✓ Media e varianza
- ✓ Funzioni di variabili aleatorie
- ✓ Distribuzioni condizionate
- ✓ Variabili aleatorie Gaussiane

Teoria della probabilità

La *teoria della probabilità* studia le proprietà globali (o di insieme) di fenomeni che si verificano sequenzialmente o contemporaneamente

Esempi: produzione di manufatti, chiamate telefoniche, traffico automobilistico, popolazioni, guasti, emissioni di elettroni, ecc.

Un *esperimento probabilistico* ha le seguenti caratteristiche:

- tutte i possibili risultati dell'esperimento sono noti a priori (*eventi elementari*)
- ogni esperimento produce un risultato che non è noto a priori
- l'esperimento può essere ripetuto nelle stesse condizioni

Un *evento* E è una collezione di eventi elementari

Classe degli eventi

Ω: insieme di tutti gli eventi elementari

S: insieme di tutti i possibili sottoinsiemi di Ω (*classe degli eventi*)

S ha le seguenti proprietà

- S è un evento
- se $E_i \in S$, allora $\bar{E}_i = S - E_i \in S$
- $\forall E_i, E_j \in S$, si ha $E_i \cap E_j \in S$ e $E_i \cup E_j \in S$

Gli elementi di S sono tutti gli eventi di cui è interessante studiare le proprietà di insieme (medie su esperimenti ripetuti, dispersione dei risultati, probabilità degli eventi, ecc.)

Legge di probabilità

Sulla classe degli eventi S è definita la *legge di probabilità* $P\{\,\}$, tale che

- $P\{E\} \geq 0$ per ogni evento E
- $P\{S\} = 1$
- se $E_1 \cap E_2 = \emptyset$, allora $P\{E_1 \cup E_2\} = P\{E_1\} + P\{E_2\}$

La terna (Ω, S, P) individua un esperimento probabilistico

Valgono le seguenti proprietà:

- $P\{\emptyset\} = 0$
- $P\{E\} = 1 - P\{\bar{E}\} \leq 1$
- $P\{E_1 \cup E_2\} = P\{E_1\} + P\{E_2\} - P\{E_1 \cap E_2\}$

Variabili aleatorie

Si dice *variabile aleatoria* una funzione dalla classe degli eventi allo spazio dei reali

$$X : S \rightarrow \mathbb{R}$$

Il valore assunto da X in corrispondenza di un evento si dice *realizzazione*

Alla variaile aleatoria è associata la *funzione di distribuzione della probabilità*

$$F_X(x) = P\{X \leq x\}$$

Proprietà:

- $P\{x_1 < x \leq x_2\} = F_X(x_2) - F_X(x_1)$
- $0 \leq F_X(x) \leq 1$
- $\lim_{x \rightarrow +\infty} F_X(x) = 1, \quad \lim_{x \rightarrow -\infty} F_X(x) = 0$
- $F_X(x)$ è monotona non decrescente: $F_X(x_1) \leq F_X(x_2) \ \ \forall x_1 \leq x_2$

Variabili aleatorie discrete e continue

Una variabile aleatoria X si dice *discreta* se esiste un insieme numerabile $\mathcal{E} \subset \mathbb{R}$ tale che $P\{X \in \mathcal{E}\} = 1$

$\Rightarrow$ La funzione di distribuzione della probabilità è costante a tratti, con discontinuità in corrispondenza degli elementi di $\mathcal{E}$

Per una v.a. discreta si può definire la funzione *massa di probabilità* come $p_i = P\{x = x_i\}$ con $p_i \geq 0$ e $\sum_i p_i = 1$. Si ha così: $F_X(x) = \sum_{x_i \leq x} p_i$

Una variabile aleatoria X si dice *continua* se la funzione di distribuzione della probabilità $F_X(x)$ è continua per ogni $x \in \mathbb{R}$

Una generica v.a. X può avere sia una componente discreta che una componente continua. In tal caso:

$$F_X(x) = F_{X,d}(x) + F_{X,c}(x)$$

Funzione densità della probabilità

Si definisce *funzione di densità della probabilità (pdf)*

$$f_X(x) = \frac{d}{dx} F_X(x)$$

Proprietà:

- $f_X(x) \geq 0$
- $\int_{-\infty}^{+\infty} f_X(x)\, dx = 1$
- $F_X(x) = \int_{-\infty}^{x} f_X(\sigma)\, d\sigma$
- $P\{a \leq X \leq b\} = \int_{a}^{b} f_X(x)\, dx$

Osservazioni:

- se X ha una componente discreta, $f_X(x)$ è definita nello spazio delle distribuzioni (contiene una delta di Dirac $\delta(x - x_i)$, per ogni $x_i \in \mathcal{E}$)
- se X è puramente discreta, $f_X(x) = \sum_i p_i \delta(x - x_i)$

Esempi di distribuzioni notevoli (continue)

Distribuzione Gaussiana (o normale)

$$f_X(x) = \frac{1}{\sqrt{2\pi}\,\sigma}\, e^{-(x-m)^2/2\sigma^2}$$

Distribuzione uniforme

$$f_X(x) = \begin{cases} \dfrac{1}{b-a} & x \in [a,b] \\ 0 & \text{altrimenti} \end{cases}$$

Distribuzione esponenziale

$$f_X(x) = \begin{cases} \dfrac{1}{\lambda}\, e^{-x/\lambda} & x \geq 0 \\ 0 & \text{altrimenti} \end{cases}$$

Esempi di distribuzioni notevoli (discrete)

Distribuzione uniforme

$$p_k = P\{X = k\} = \frac{1}{n} \qquad k = 1, \ldots, n$$

Distribuzione binomiale

$$p_k = P\{X = k\} = \binom{n}{k} p^k (1-p)^{n-k} \qquad k = 0, 1, \ldots, n \quad 0 \leq p \leq 1$$

Distribuzione geometrica

$$p_k = P\{X = k\} = p(1-p)^k \qquad k = 0, 1, \ldots \qquad 0 \leq p \leq 1$$

Distribuzione di Poisson

$$p_k = P\{X = k\} = e^{-\lambda} \frac{\lambda^k}{k!} \qquad k = 0, 1, \ldots \qquad \lambda > 0$$

Distribuzioni bivariate

X, Y: variabili aleatorie

- $F_{X,Y}(x,y) = P\{X \leq x; Y \leq y\}$ distribuzione bivariata
- $f_{X,Y}(x,y) = \dfrac{d^2}{dx\,dy} F_{X,Y}(x,y)$ densità di probabilità bivariata

Proprietà:

- $F_{X,Y}(x,y)$ monotona non decrescente in x e y
- $\lim\limits_{x,y\to+\infty} F_{X,Y}(x,y) = 1, \quad \lim\limits_{x,y\to-\infty} F_{X,Y}(x,y) = 0$
- $P\{(X,Y) \in \mathcal{A}\} = \iint_{\mathcal{A}} f_{X,Y}(x,y)\,dx\,dy$

X e Y sono variabili aleatorie **indipendenti** se

$$f_{X,Y}(x,y) = f_X(x)\, f_Y(y)$$

Distribuzioni multivariate

$X \in \mathbb{R}^n$: vettore di n variabili aleatorie $\qquad X = [X_1 \; X_2 \; \ldots \; X_n]'$

$f_X(x)$: densità di probabilità multivariata

Proprietà:

- $\int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \ldots \int_{-\infty}^{+\infty} f_X(x)\, dx_n dx_{n-1} \ldots dx_1 \;=\; 1$
- $f_{X_i}(x_i) \;=\; \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \ldots \int_{-\infty}^{+\infty} f_X(x)\, dx_n \ldots dx_{i+1} dx_{i-1} \ldots dx_1$

$X_1, X_2, \ldots, X_n$ sono variabili aleatorie **indipendenti** se

$$f_X(x) = f_{X_1}(x_1)\, f_{X_2}(x_2) \; \ldots \; f_{X_n}(x_n)$$

Operatore di media

Si definisce **media** (o **valore atteso**) di una variabile aleatoria X

$$E[X] = \int_{-\infty}^{+\infty} x\, f_X(x)\, dx \triangleq m_X$$

e di una generica funzione $g(X)$

$$E[g(X)] = \int_{-\infty}^{+\infty} g(x)\, f_X(x)\, dx$$

Ponendo $g(X) = X^p$ si ottengono i *momenti di ordine p*.

Proprietà: l'operatore di media è lineare

$$E[\alpha X + \beta Y] = \alpha E[X] + \beta E[Y]$$

Varianza e covarianza

Si definisce **varianza** di una variabile aleatoria X

$$\mathrm{Var}(X) \;=\; E[(X - m_X)^2] \;=\; \int_{-\infty}^{+\infty} (x - m_X)^2 \, f_X(x)\, dx \;\triangleq\; \sigma_X^2$$

σ_X: deviazione standard di X

Si definisce **covarianza** di due variabili aleatorie X, Y

$$E[(X - m_X)(Y - m_Y)] = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} (x - m_X)(y - m_Y)\, f_{X,Y}(x, y)\, dydx \;\triangleq\; P_{XY}$$

Caso vettoriale ($X \in \mathbb{R}^n$, $Y \in \mathbb{R}^m$)

- $P_X = E[(X - m_X)(X - m_X)'] \in \mathbb{R}^{n \times n}$
- $P_{XY} = E[(X - m_X)(Y - m_Y)'] \in \mathbb{R}^{n \times m}$
- X e Y si dicono **scorrelate** se $P_{XY} = 0$
- X e Y indipendenti $\Rightarrow$ X e Y scorrelate

Variabili aleatorie Gaussiane

$X \in \mathbb{R}^n$ sono variabili aleatorie **congiuntamente Gaussiane** se la loro pdf è

$$f_X(x) = \frac{1}{\sqrt{(2\pi)^n \det(P_X)}} e^{-\frac{1}{2}(x-m_X)' P_X^{-1} (x-m_X)}$$

con $P_X = P_X' > 0$.

Notazione: $X \sim \mathcal{N}(m_X, P_X)$

Proprietà:

- $E[X] = m_X$
- $E[(X - m_X)(X - m_X)'] = P_X$
- X cong. Gaussiane $\Rightarrow Y = AX + b$ cong. Gaussiane, con $m_Y = Am_X + b, \quad P_Y = AP_XA'$
- X e Y v.a. scorrelate e cong. Gaussiane $\Rightarrow X$ e Y indipendenti

Teorema del limite centrale

Siano $X_j \in \mathbb{R}$, $j = 1, 2, \ldots,$ variabili aleatorie indipendenti e identicamente distribuite, tali che $E[X_j] = m_X$ e $\mathrm{Var}(X_j) = \sigma_X^2$. Allora, per ogni $x \in \mathbb{R}$

$$\lim_{n\to\infty} P\left\{ \frac{\bar{X} - m_X}{\sigma_X/\sqrt{n}} \leq x \right\} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-u^2/2} du$$

$\Rrightarrow$ La media campionaria $\bar{X}$ di variabili aleatorie equidistribuite, con media m_X e varianza σ_X^2, è asintoticamente normalmente distribuita, con media m_X e varianza σ_X^2/n, ovvero

$$\lim_{n\to\infty} \bar{X} \sim \mathcal{N}\left(m_X, \frac{\sigma_X^2}{n} \right)$$

Funzioni di variabili aleatorie

Sia $X \in \mathbb{R}$ v.a. con pdf $f_X(x)$. Sia $Y \in \mathbb{R}$ v.a., definita da $Y = g(X)$.
Si ha

$$f_Y(y) = \sum_{i=1}^{m} \frac{f_X(x_i)}{|g'(x_i)|}$$

dove $g'(x) = \dfrac{dg(x)}{dx}$ e $g(x_1) = g(x_2) = \cdots = g(x_m) = y$.

Caso vettoriale ($X, Y \in \mathbb{R}^n$):

$$f_Y(y) = \sum_{i=1}^{m} \frac{f_X(x_i)}{|J(x_i)|}$$

dove $J(x_i) = \det\left(\dfrac{\partial g(x)}{\partial x}\right)\Big|_{x=x_i}$ e $g(x_1) = g(x_2) = \cdots = g(x_m) = y$.

Probabilità condizionata

Dati due eventi A e B, la probabilità dell'evento A *condizionata* all'evento B è data da

$$P\{A|B\} = \frac{P\{A \cap B\}}{P\{B\}}$$

La funzione di distribuzione della probabilità di una v.a. X, *condizionata* all'evento $Y = y$, è definita da

$$F_{X|Y}(x|y) = \lim_{\varepsilon \to 0+} P\{X \leq x \mid \ y - \varepsilon \leq Y \leq y + \varepsilon\}$$

La funzione di *densità di probabilita condizionata* di X rispetto a Y è

$$f_{X|Y}(x|y) = \frac{d}{dx} F_{X|Y}(x|y)$$

Distribuzioni condizionate

Formula di Bayes

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}$$

Conseguenze:

$$\Rightarrow \quad f_X(x) = \int_{-\infty}^{+\infty} f_{X|Y}(x|y)\, f_Y(y)\, dy$$

$\Rightarrow$ Se X e Y sono indipendenti: $f_{X|Y}(x|y) = f_X(x)$

Definizioni:

- **media condizionata**: $E[X|Y] = \int_{-\infty}^{+\infty} x\, f_{X|Y}(x|y)\, dx$
- **varianza condizionata**: $P_{X|Y} = \int_{-\infty}^{+\infty} (x - E[X|Y])^2\, f_{X|Y}(x|y)\, dx$

Distribuzione Gaussiana condizionata

Siano X e Y v.a. congiuntamente Gaussiane tali che:

$$E[X] = m_X \qquad E[Y] = m_Y$$

$$E\left[\begin{pmatrix} X - m_X \\ Y - m_Y \end{pmatrix}\begin{pmatrix} X - m_X \\ Y - m_Y \end{pmatrix}'\right] = \begin{pmatrix} R_X & R_{XY} \\ R'_{XY} & R_Y \end{pmatrix}$$

Risultato:

$$\begin{aligned} E[X|Y] &= m_X + R_{XY}R_Y^{-1}(Y - m_Y) \\ P_{X|Y} &= R_X - R_{XY}R_Y^{-1}R'_{XY} \end{aligned}$$

Processi stocastici

Si dice *processo stocastico* (p.s.), una sequenza temporale di variabili aleatorie $x(t)$, con $t \in \mathcal{T}$, dove $\mathcal{T}$ rappresenta l'asse dei tempi.

- se $\mathcal{T}$ è un insieme numerabile $\{t_1, t_2, \ldots, t_k, \ldots\}$, il processo stocastico è detto *tempo-discreto*
- se $\mathcal{T} = \{t \in \mathbb{R}, t \geq 0\}$, il processo stocastico è detto *tempo-continuo*

Dalla definizione di variabile aleatoria, un p.s. è una funzione $x(t, \omega)$:

$$x : \mathcal{T} \times \Omega \rightarrow \mathbb{R}.$$

$\rightarrow$ fissato un istante temporale $\bar{t} \in \mathcal{T}$, $x(\bar{t})$ è una variabile aleatoria

$\rightarrow$ fissato l'evento ω, $x(t)$ è una funzione deterministica di t (detta *realizzazione* del processo stocastico)

Statistiche di un processo stocastico

Un processo stocastico è completamente caratterizzato in senso statistico se è nota la probabilità:

$$P\{x(t_1) \leq x_1, x(t_2) \leq x_2, \ldots, x(t_k) \leq x_k\}, \quad \forall t_1, \ldots, t_k \quad \forall x_1, \ldots, x_k \quad \forall k.$$

Analogamente alle variabili aleatorie, è possibile definire le funzioni di *distribuzione* e di *densità* della probabilità per ogni $k = 1, 2, \ldots$ (*statistiche di ordine k*):

- distribuzione

$$F_x(x_1, \ldots, x_k; t_1, \ldots, t_k)$$

- densità

$$f_x(x_1, \ldots, x_k; t_1, \ldots, t_k)$$

Esempi:

- statistiche del primo ordine

$$\begin{cases} F_x(x;t) = P\{x(t) \leq x\} \\ f_x(x;t) = \dfrac{\partial}{\partial x} F_x(x;t) \end{cases}$$

- statistiche del secondo ordine

$$\begin{cases} F_{x_1,x_2}(x_1,x_2;t_1,t_2) = P\{x(t_1) \leq x_1, x(t_2) \leq x_2\} \\ f_{x_1,x_2}(x_1,x_2;t_1,t_2) = \dfrac{\partial^2}{\partial x_1 \partial x_2} F_{x_1,x_2}(x_1,x_2;t_1,t_2) \end{cases}$$

Media e covarianza di processi stocastici

Si definisce **media** (o **valore atteso**) $m_{\boldsymbol{x}}(t)$ del p.s. $x(t)$

$$m_{\boldsymbol{x}}(t) \;=\; E[x(t)] \;=\; \int_{-\infty}^{+\infty} x\, f_x(x;t)\,\mathrm{d}x$$

Si definisce **funzione di covarianza** (o **autocovarianza**) $R_{m_{\boldsymbol{x}}}(t,s)$ di un p.s. $x(t)$

$$\begin{aligned} R_{\boldsymbol{x}}(t,s) &= E[(x(t)-m_{\boldsymbol{x}}(t))(x(s)-m_{\boldsymbol{x}}(s))'] \\ &= \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} (x_1-m_{\boldsymbol{x}}(t))(x_2-m_{\boldsymbol{x}}(s))'\, f_{x_1,x_2}(x_1,x_2;t,s)\,\mathrm{d}x_1\mathrm{d}x_2 \end{aligned}$$

Si definisce **funzione di covarianza incrociata** (o **cross-covarianza**) $R_{\boldsymbol{x},\boldsymbol{y}}(t,s)$ di due p.s. $x(t)$, $y(t)$

$$\begin{aligned} R_{\boldsymbol{x},\boldsymbol{y}}(t,s) &= E[(x(t)-m_{\boldsymbol{x}}(t))(y(s)-m_{\boldsymbol{y}}(s))'] \\ &= \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} (x-m_{\boldsymbol{x}}(t))(y-m_{\boldsymbol{y}}(s))'\, f_{x,y}(x,y;t,s)\,\mathrm{d}x\mathrm{d}y \end{aligned}$$

Caso scalare: $x, y \in \mathbb{R}$

- $m_{\boldsymbol{x}}(t) : \mathbb{R} \to \mathbb{R}$
- $R_{\boldsymbol{x}}(t,s) : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$
- $R_{\boldsymbol{x},\boldsymbol{y}}(t,s) : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$

Caso vettoriale: $x \in \mathbb{R}^n$, $y \in \mathbb{R}^m$

- $m_{\boldsymbol{x}}(t) : \mathbb{R} \to \mathbb{R}^n, \qquad m_{\boldsymbol{y}}(t) : \mathbb{R} \to \mathbb{R}^m$
- $R_{\boldsymbol{x}}(t,s) : \mathbb{R} \times \mathbb{R} \to \mathbb{R}^{n \times n}, \qquad R_{\boldsymbol{y}}(t,s) : \mathbb{R} \times \mathbb{R} \to \mathbb{R}^{m \times m}$
- $R_{\boldsymbol{x},\boldsymbol{y}}(t,s) : \mathbb{R} \times \mathbb{R} \to \mathbb{R}^{n \times m}$

Stazionarietà

Un processo stocastico $x(t)$ si dice **stazionario in senso forte** se le statistiche di ogni ordine di $x(t)$ e $x(t+\tau)$ sono identiche, ovvero se:

$$P\{x(t_1+\tau) \leq x_1, \ldots, x(t_k+\tau) \leq x_k\}$$

non dipende da τ, $\forall t_1, \ldots, t_k$, $\forall x_1, \ldots, x_k$ e $\forall k$.

Un processo stocastico $x(t)$ si dice **stazionario in senso debole** se i momenti del primo e del secondo ordine di $x(t)$ e $x(t+\tau)$ sono identici, ovvero se per ogni τ

$$m_{\boldsymbol{x}}(t) = m_{\boldsymbol{x}}(t+\tau)$$
$$R_{\boldsymbol{x}}(t,s) = R_{\boldsymbol{x}}(t+\tau, s+\tau)$$

$\Rightarrow$ un p.s. $x(t)$ stazionario in senso debole se e solo se

$$m_{\boldsymbol{x}}(t) = m_{\boldsymbol{x}} \qquad R_{\boldsymbol{x}}(t,s) = R_{\boldsymbol{x}}(t-s)$$

$\Rightarrow$ la stazionarietà forte implica quella debole (ma non viceversa)

Processi gaussiani, processi bianchi

Un processo stocastico si dice **gaussiano** se le sue statistiche:

$$F_x(x_1, \ldots, x_k; t_1, \ldots, t_k)$$

sono gaussiane, per ogni k.

Proprietà:

- un p.s. gaussiano è caratterizzato completamente da un punto di vista statistico qualora siano note la media e la funzione di covarianza
- per i p.s. gaussiani la stazionarietà forte e quella debole coincidono

Un processo stocastico $x(t)$ si dice **bianco** se $x(t_1)$ e $x(t_2)$ sono variabili aleatorie indipendenti, per ogni coppia $t_1 \neq t_2$.
Se le variabili aleatorie sono anche identicamente distribuite, si parla di processo a variabili $i.i.d.$ (indipendenti e identicamente distribuite).

Lezione # 2

- ✓ MATLAB Statistics Toolbox
- ✓ Generazione di numeri casuali
- ✓ Caratteristiche delle principali distribuzioni
- ✓ Statistiche descrittive
- ✓ Plot statistici

MATLAB Statistics Toolbox - Version 3.0

Un insieme di strumenti di calcolo per problemi di statistica e calcolo delle probabilità

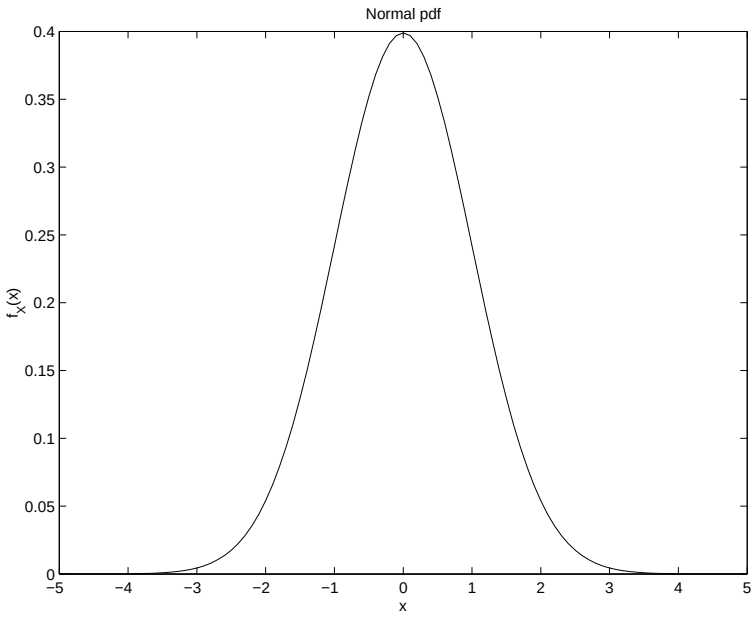
Oltre 200 funzioni, divise in vari gruppi

- Probability density functions (pdf)
- Cumulative Distribution functions (cdf)
- Random Number Generators
- Statistics
- Linear and Nonlinear Models
- Parameter estimation
- Multivariate Statistics
- Hypothesis Tests
- Statistical Plotting
- etc...

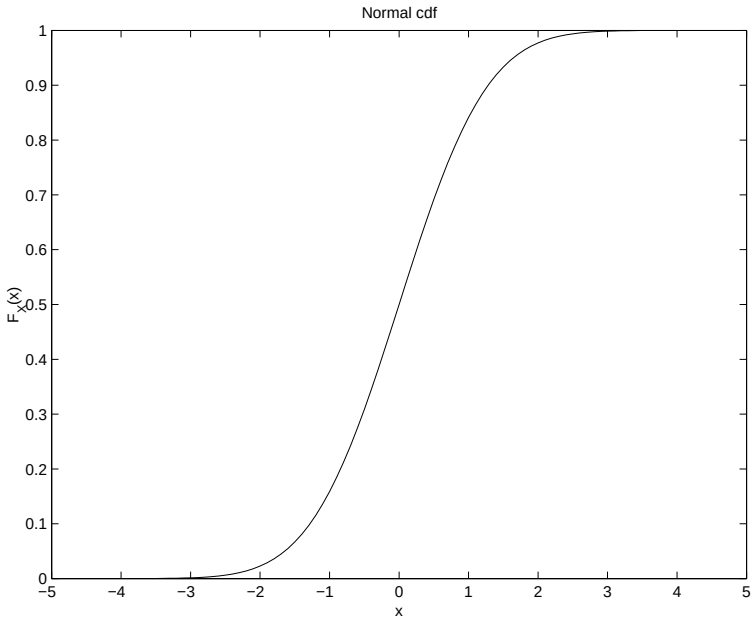
Elenco completo → `>> help stats`

Comandi pdf e cdf

```
>> x=[-5:0.1:5];
>> f=normpdf(x,0,1); % pdf Gaussiana, media=0, varianza=1
>> F=normcdf(x,0,1); % cdf Gaussiana, media=0, varianza=1
>> figure, plot(x,f)
>> figure, plot(x,F)
```



$f_X(x)$ Gaussiana



$F_X(x)$ Gaussiana

Generatori di numeri casuali

→ uso dei comandi **rnd** (20 distribuzioni predefinite)

```
>> y=normrnd(0,1,1000,1);
>> % vettore 1000 × 1 di v.a. Gaussiane (media=0, varianza=1)
```

→ metodo di inversione

Se u è una v.a. distribuita uniformemente tra 0 e 1, $F^{-1}(u)$ è una v.a. con distribuzione F

```
>> u=rand(1000,1);
>> y=-log(1-u);
>> % vettore 1000 × 1 di v.a. con distribuzione esponenziale (λ = 1)
```

Esercizio: Generare 1000 v.a. distribuite con pdf

$$f_X(x) = \begin{cases} 1.5x^2 & x \in [-1, 1] \\ 0 & \text{altrimenti} \end{cases}$$

Media e varianza

Per distribuzioni assegnate, media e varianza si ottengono utilizzando i comandi `stat`

```
>> [M,V]=expstat(1); % pdf esponenziale (lambda=1)
>> [M,V]=unifstat(-1,1); % pdf uniforme in [-1,1]
```

Media e varianza campionaria di un campione di dimensione n si ottengono tramite i comandi `mean` e `cov`

```
>> y=exprnd(1,1000,1);
>> Mc=mean(y); % media campionaria
>> Vc=cov(y); % varianza campionaria
```

Statistiche campionarie

Siano $X_1, X_2, \ldots, X_n$, n osservazioni indipendenti di una variabile aleatoria X, con distribuzione F_X. L'insieme X_i, $i = 1, \ldots, n$, è un *campione* di dimensione n della distribuzione F_X.

Una funzione $h(X_1, X_2, \ldots, X_n)$ è una *statistica campionaria* della v.a. X

- Media campionaria

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

  Proprietà: $E[\bar{X}] = m_X$, $\quad \mathrm{Var}(\bar{X}) = E[(\bar{X} - m_X)^2] = \dfrac{\sigma_X^2}{n}$

- Varianza campionaria

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

  Proprietà: $E[S_X^2] = \sigma_X^2$

Esempi di distribuzioni

Distribuzioni discrete

- ✓ Distribuzione binomiale
- ✓ Distribuzione geometrica
- ✓ Distribuzione di Poisson

Distribuzioni continue

- ✓ Distribuzione uniforme (continua)
- ✓ Distribuzione log-normale
- ✓ Distribuzione Gamma
- ✓ Distribuzione esponenziale
- ✓ Distribuzione χ^2

Distribuzione binomiale

Modella il numero di esiti positivi su n prove nelle ipotesi seguenti:

- solo due risultati possibili per ogni prova (positivo o negativo)
- la probabilità di esito positivo per ogni prova è p, $0 \leq p \leq 1$
- le prove sono indipendenti

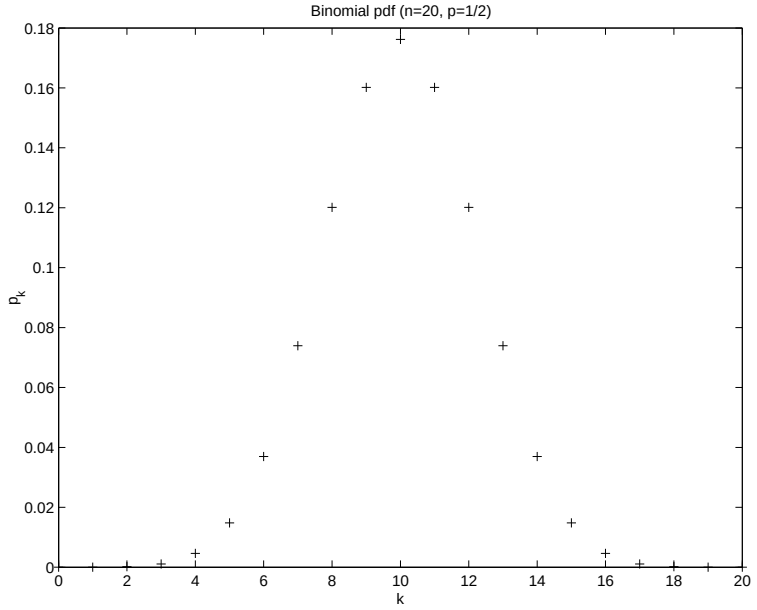
La probabilità di avere k esiti positivi è pari a

$$p_k = P\{X = k\} = \binom{n}{k} p^k (1-p)^{n-k} \qquad k = 0, 1, \ldots, n$$

Proprietà:

$$E[X] = np; \quad \mathrm{Var}(X) = np(1-p)$$

```
>> n=20;
>> p=1/2;
>> x=0:n;
>> y=binopdf(x,n,p);
>> plot(x,y,'+')
```



Distribuzione geometrica

Modella il numero di esiti negativi consecutivi, prima che si verifichi un esito positivo (stesse ipotesi viste per la distribuzione binomiale)

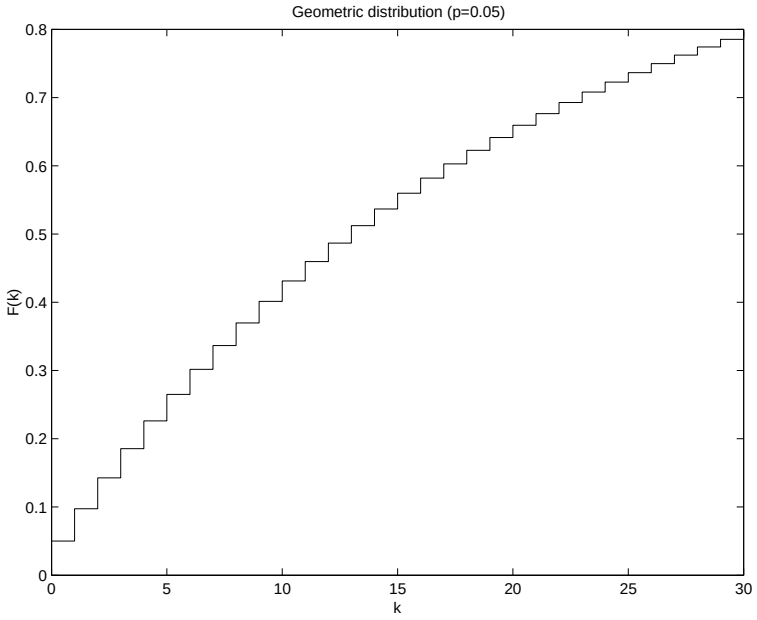
La probabilità di avere k esiti negativi consecutivi prima di un esito positivo è pari a

$p_k = P\{X = k\} = p(1-p)^k \qquad k = 0, 1, \ldots$

Proprietà:

$E[X] = \frac{1-p}{p}; \quad \mathrm{Var}(X) = \frac{1-p}{p^2}$

```
>> n=30;
>> p=0.05;
>> x=0:n;
>> y=geocdf(x,p);
>> stairs(x,y)
```



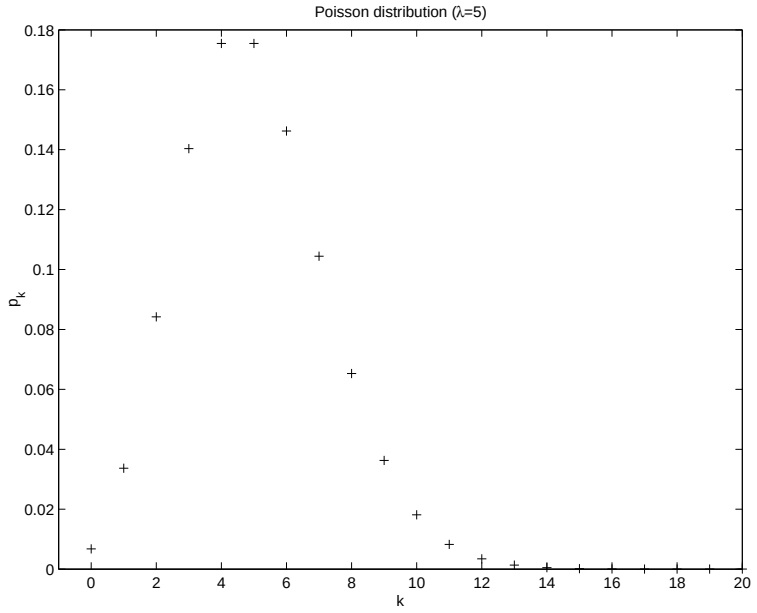
Distribuzione di Poisson

Modella il numero di volte che un certo evento si verifica in una unità di tempo o di spazio
Esempi: numero di clienti di un negozio in un giorno; numero di auto che transitano in un'ora; ecc.

La probabilità di ottenere k eventi è pari a
$p_k = P\{X = k\} = e^{-\lambda}\frac{\lambda^k}{k!} \qquad k = 0, 1, \ldots \text{ con } \lambda > 0$

Proprietà:
$E[X] = \text{Var}(X) = \lambda$

```
>> n=20;
>> lambda=5;
>> x=0:n;
>> y=poisspdf(x,lambda);
>> plot(x,y,'+')
```



Distribuzione uniforme (continua)

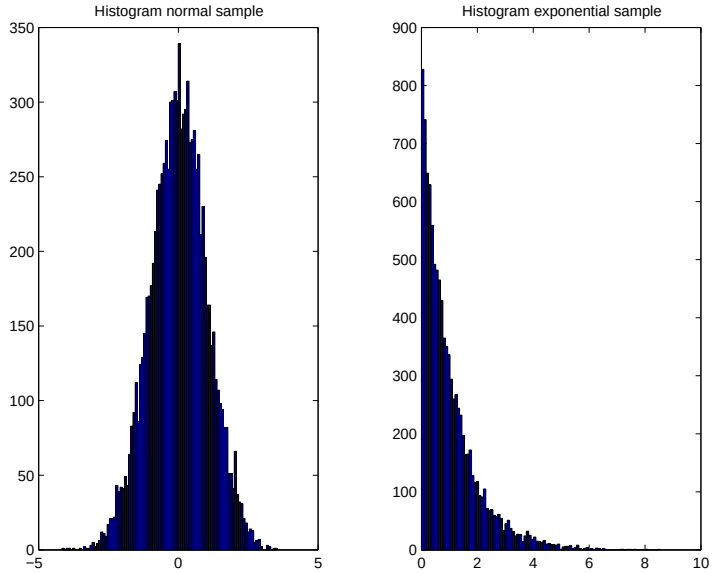
$$f_X(x;a,b) = \begin{cases} \dfrac{1}{b-a} & x \in [a,b] \\ 0 & \text{altrimenti} \end{cases}$$

Nel caso $a = 0$, $b = 1$: distribuzione uniforme *standard*
(comandi MATLAB: `rand`, `unifrnd(a,b)`)

Esempi di applicazioni: distribuzione degli errori di arrotondamento; generazione di numeri casuali da distribuzioni assegnate (inversione)

Proprietà:
$E[X] = \frac{a+b}{2}$; $\mathrm{Var}(X) = \frac{(b-a)^2}{12}$

```
>> u=rand(10000,1);
>> x=norminv(u,0,1);
>> y=expinv(u,1);
>> subplot(121), hist(x,100);
>> subplot(122), hist(y,100);
```



Distribuzione log-normale

$$f_X(x; m, \sigma) = \frac{1}{\sqrt{2\pi}\,\sigma\,x}\, e^{-(\ln x - m)^2/2\sigma^2} \qquad x > 0$$

Se X è una v.a. con distribuzione log-normale con parametri m e σ, allora $\ln X$ ha distribuzione normale con media m e varianza σ^2

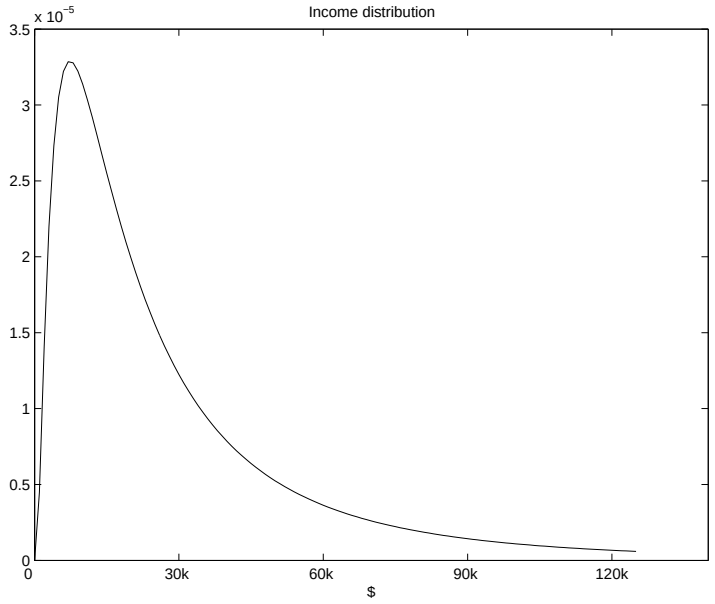
Applicazioni in campo economico (distribuzione del reddito, ecc.)

Proprietà:

$E[X] = \exp\left(m + \frac{\sigma^2}{2}\right)$;

$\mathrm{Var}(X) = \exp(2m + 2\sigma^2) - \exp(2m + \sigma^2)$

```
>> x=[0:1000:125000];
>> y=lognpdf(x,log(20000),1);
>> plot(x,y)
>> title('Income distribution')
```



Distribuzione Gamma

$$f_X(x;a,b) = \begin{cases} \dfrac{1}{b^a \Gamma(a)} x^{a-1} e^{-x/b} & x > 0 \\ 0 & \text{altrimenti} \end{cases}$$

dove $\Gamma(a) = \int_0^\infty y^{a-1} e^{-y} dy$ è la *funzione Gamma*

Distribuzione Gamma a due parametri a e b (`gampdf(x,a,b)`)

Applicazioni: modella la vita media dei componenti elettronici, la durata media del funzionamento corretto di dispositivi, ecc.

Casi particolari:

- $a = 1$, $b = \lambda$: distribuzione *esponenziale* con parametro λ
- $b = 2$, $a = n/2$, n intero positivo: distribuzione χ^2 a n gradi di libertà ($\chi^2(n)$)

Distribuzione esponenziale

$$f_X(x;\lambda) = \begin{cases} \frac{1}{\lambda}\, e^{-x/\lambda} & x \geq 0 \\ 0 & \text{altrimenti} \end{cases}$$

Proprietà:

- $E[X] = \lambda$; $\mathrm{Var}(X) = \lambda^2$
- $\dfrac{F_X(x_2) - F_X(x_1)}{1 - F_X(x_1)} = F_X(x_2 - x_1)$

  (probabilità di sopravvivenza per una ulteriore unità di tempo)

```
>> life=10:10:60;
>> lifeplus1=l+1;
>> deltap=(expcdf(lifeplus1,10)-expcdf(life,10))./(1-expcdf(life,10))

deltap =

   0.0952    0.0952    0.0952    0.0952    0.0952    0.0952
```

Distribuzione χ^2

$$f_X(x;n) = \begin{cases} \dfrac{1}{\Gamma(n/2)2^{n/2}} e^{-x/2} x^{n/2-1} & x > 0 \\ 0 & \text{altrimenti} \end{cases}$$

Proprietà:

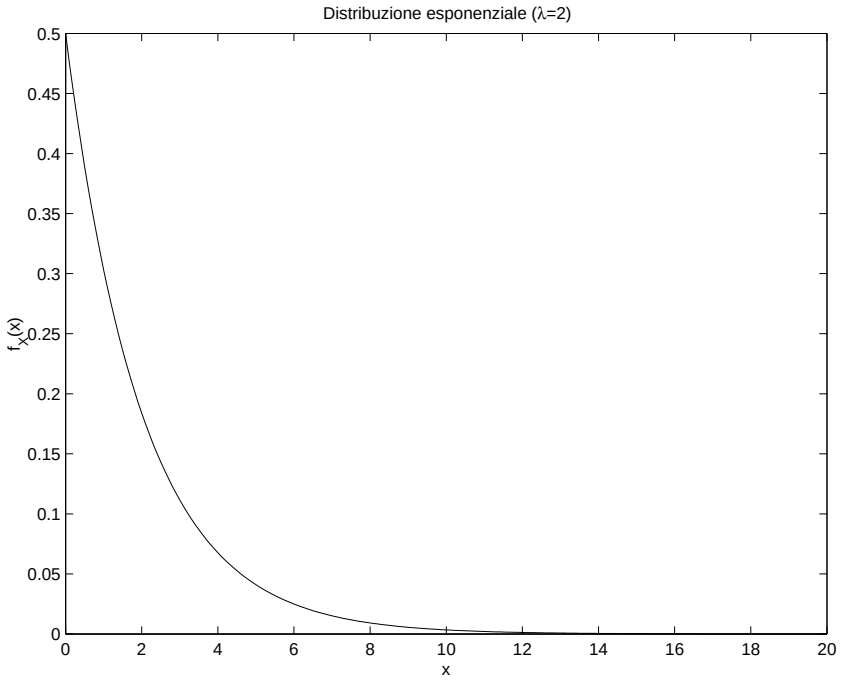
- $E[X] = n$; $\mathrm{Var}(X) = 2n$
- Date n variabili aleatorie $X_i \sim \mathcal{N}(m_X, \sigma_X^2)$, allora

$$\sum_{i=1}^{n} \frac{(X_i - m_X)^2}{\sigma_X^2} \sim \chi^2(n)$$

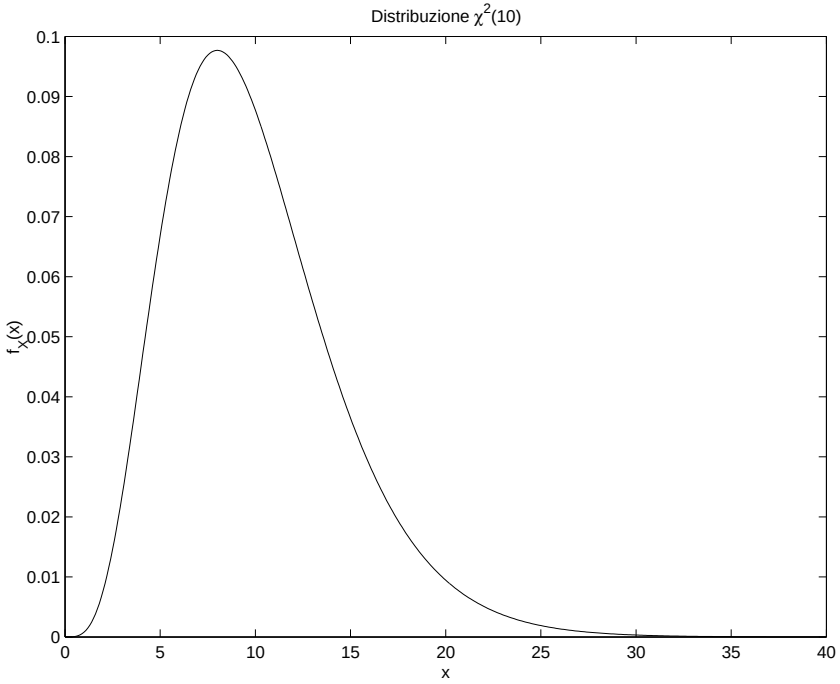
- Date n osservazioni indipendenti di una variabile aleatoria $X \sim \mathcal{N}(m_X, \sigma_X^2)$, la varianza campionaria S_X^2 è tale che

$$\frac{(n-1)S_X^2}{\sigma_X^2} \sim \chi^2(n-1)$$

Distribuzione esponenziale e distribuzione χ^2



$f_X(x)$ esponenziale $(\lambda = 2)$



$f_X(x)$ $\chi^2(10)$

Statistiche descrittive

Le *statistiche descrittive* forniscono indicatori numerici che riassumono le caratteristiche di un campione di dati

→ Sono utili specialmente per campioni molto numerosi

Esempi:

- Statistiche campionarie
- Misure di locazione
- Misure di dispersione
- Percentili

Percentili

Sia X una variabile aleatoria.
Un numero x che soddisfa

$$P\{X \leq x\} \geq p \qquad P\{X \geq x\} \geq 1-p \qquad 0 < p < 1$$

si dice *quantile di ordine* p, o *percentile al* $100p\,\%$ della v.a. X.

Per una v.a. continua, il quantile di ordine p si ottiene risolvendo l'equazione

$$F_X(x) = p$$

Se la v.a. contiene una componente discreta per $X = x$, si ha:
$p \leq F_X(x) \leq p + P\{X = x\}$

Il percentile al $50\,\%$ si dice *mediana* di X

Statistiche descrittive: misure di locazione

X_i, $i = 1, \ldots, n$: campione

- Media campionaria $\bar{X}$ (`mean`)
- Mediana campionaria (`median`)

$$X_{\text{med}} \ : \ \#\{X_i \leq X_{\text{med}}\} \ = \ \#\{X_i \geq X_{\text{med}}\}$$

- Media geometrica (`geomean`)

$$X_{\text{geom}} = \sqrt[n]{\prod_{i=1}^{n} X_i}$$

- Media armonica (`harmmean`)

$$X_{\text{arm}} = \frac{1}{\frac{1}{n}\sum_{i=1}^{n} \frac{1}{X_i}}$$

Statistiche descrittive: misure di dispersione

X_i, $i = 1, \ldots, n$: campione

- Varianza campionaria S_X^2 (`var`, `cov`), Deviazione standard S_X (`std`)
- Range (`range`)

$$\text{Range} = \max_i X_i - \min_i X_i$$

- Deviazione media in modulo (`mad`)

$$\text{MAD} = \frac{1}{n} \sum_{i=1}^{n} |X_i - \bar{X}|$$

- Range tra percentili fissati (`iqr`)

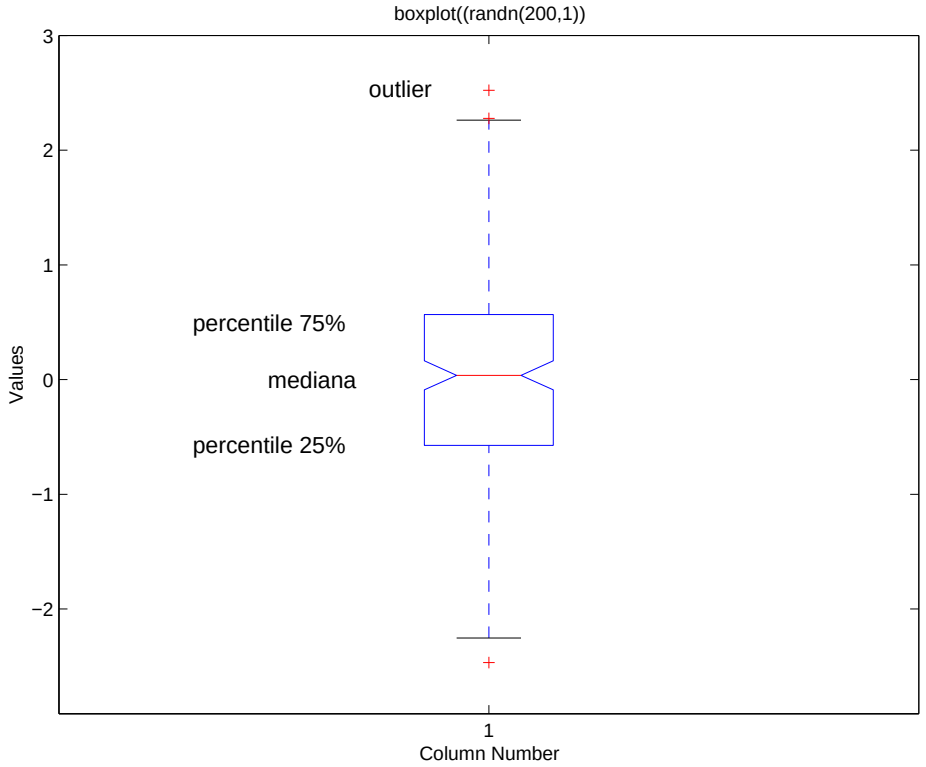
$$\text{IQR} = X_{p,75\%} - X_{p,25\%}$$

dove $X_{p,r\%}$ è tale che $\dfrac{\#\{X_i \leq X_{p,r\%}\}}{n} \times 100 = r$
(percentile a $r\,\%$, `prctile`)

Plot statistici

Rappresentazione grafica di campioni di dati (`boxplot`)

```
>> y=randn(200,1);
>> boxplot(y,1)
```



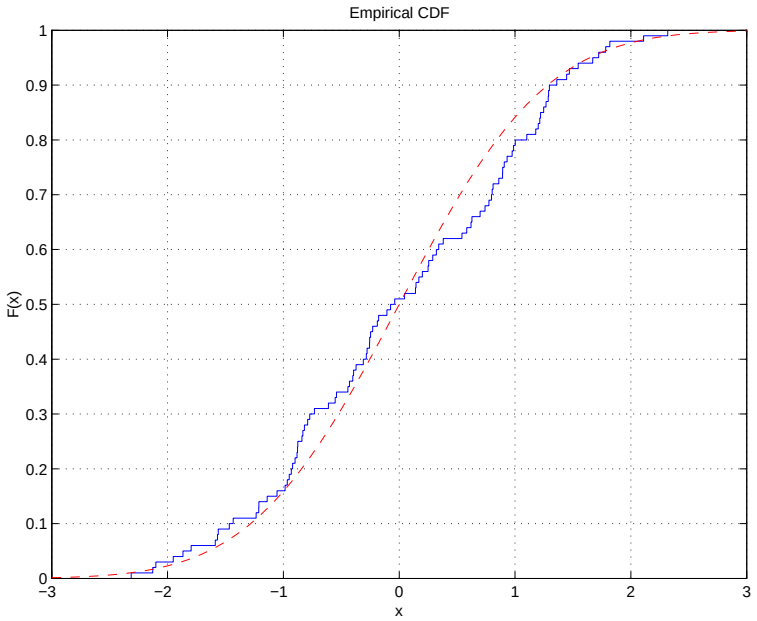
Distribuzione empirica

Dato un campione X_i, $i = 1, \ldots, n$, la sua *distribuzione empirica* è definita da

$$F_X^*(x) = \frac{1}{n} \sum_{j=1}^{n} \varepsilon(x - X_j)$$

dove $\varepsilon(x) = \begin{cases} 1 & x \geq 0 \\ 0 & x < 0 \end{cases}$.

```
>> y=randn(100,1);
>> cdfplot(y)
>> x=[-3:0.1:3];
>> yt=normcdf(x,0,1);
>> hold on; plot(x,yt,'r--')
```

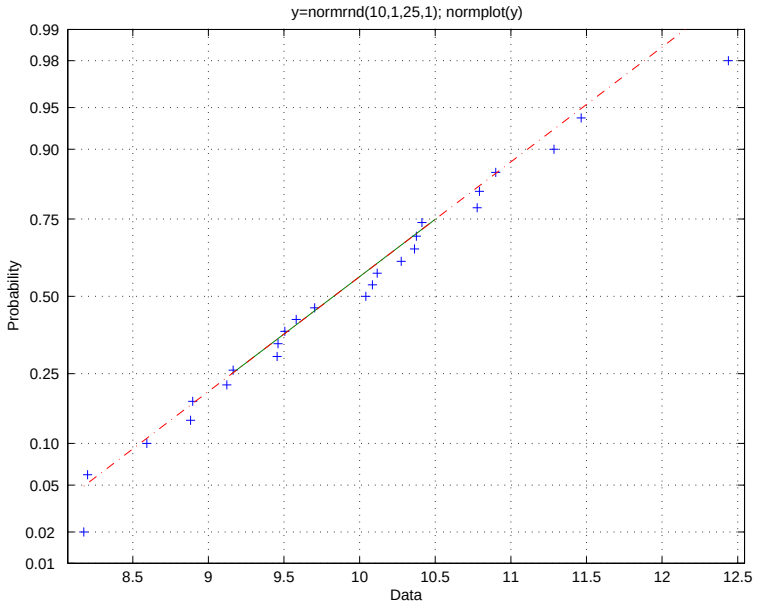


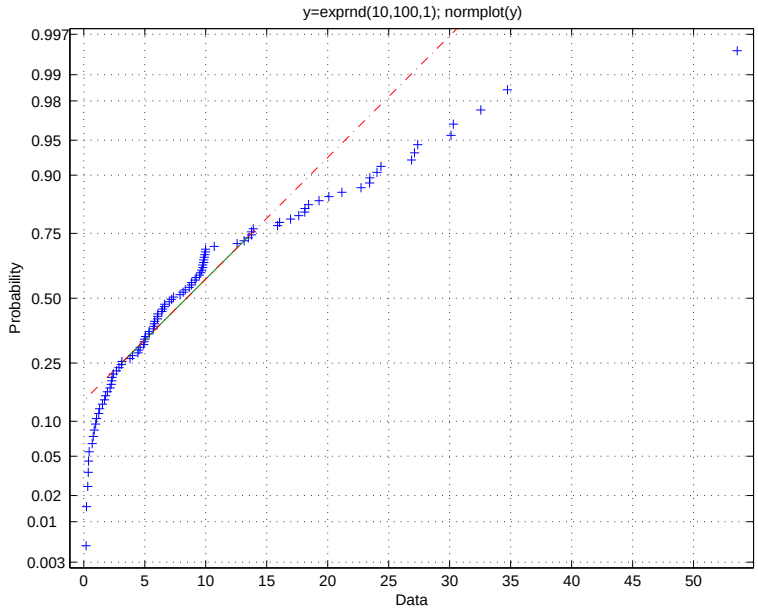
Distribution plots

Sono utili per stabilire se un campione può essere stato generato secondo una certa distribuzione o per confrontare tra loro campioni diversi

Esempi: normal probability plot, quantile-quantile plot, etc.

```
>> y=normrnd(10,1,25,1); normplot(y)
>> z=exprnd(10,100,1); normplot(z)
```





Scatter plots

Sono grafici che rappresentano un campione in funzione di un altro, con la possibilità di raggruppare i campioni in funzione di una terza variabile

```
>> load carsmall
>> gscatter(Weight,MPG,Model_Year,'','xos')
>> xvars=[Weight Horsepower]; yvars=[MPG Acceleration];
>> gplotmatrix(xvars,yvars,Model_Year,'','xos')
```

