Variabili aleatorie

Statistica e Elementi di Probabilità

Bilotti Alessandro
matricola: 206409

Variabili Aleatorie Discrete

Una **variabile aleatoria discreta** è una variabile che può assumere un **insieme finito o numerabile di valori**, ciascuno con una probabilità specifica.

Esempio: Lancio di un dado

Consideriamo una variabile aleatoria X che rappresenta il risultato del lancio di un dado a sei facce.
I possibili valori di X sono:

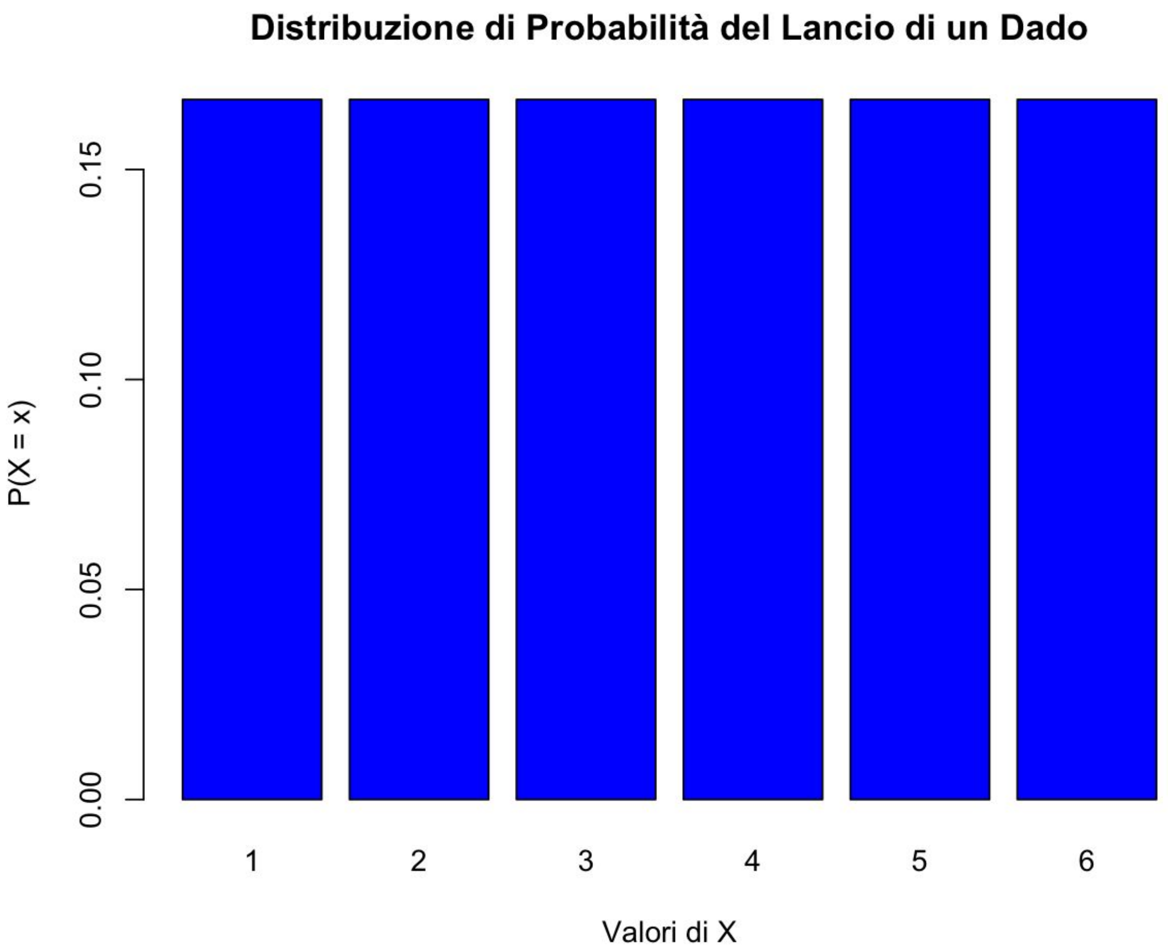
$$X \in \{1, 2, 3, 4, 5, 6\}$$

Dato che il dado è equo, la probabilità di ciascun valore è:

$$P(X = x) = \frac{1}{6}$$

Con funzione di **massa di probabilità (PMF)** data da:

$$P(X = x) = \begin{cases} \frac{1}{6}, & x \in \{1, 2, 3, 4, 5, 6\} \\ 0, & \textit{altrimenti} \end{cases}$$


Distribuzione di Probabilità del Lancio di un Dado
P(X = x)
0.00
0.05
0.10
0.15
1
2
3
4
5
6
Valori di X

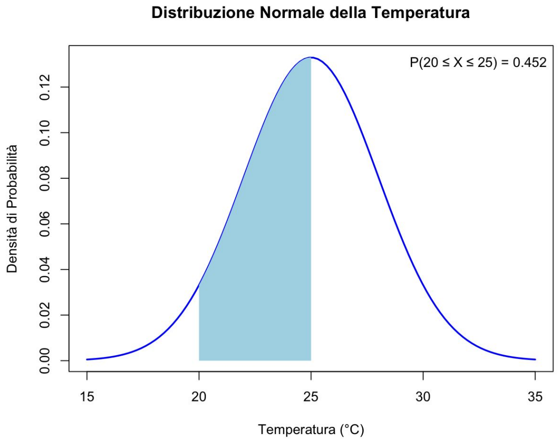
Variabili Aleatorie Continue

Una **variabile aleatoria continua** può assumere **infiniti valori** in un intervallo della retta reale. La sua distribuzione è descritta da una **funzione di densità di probabilità (PDF)** anziché da una funzione di **massa di probabilità (PMF)**.

Esempio: Temperatura giornaliera

Consideriamo una variabile aleatoria X che rappresenta la temperatura giornaliera in una città.

- X può assumere qualsiasi valore reale in un intervallo (es. $X \in [15° C,\ 35° C]$);
- È impossibile assegnare una probabilità a un singolo valore, ma è possibile calcolare la probabilità che X sia in un certo intervallo;
- La distribuzione della temperatura può essere modellata con una **distribuzione normale** $X \sim \mathcal{N}(\mu, \sigma^2)$.



Funzione di ripartizione

La **funzione di ripartizione** o **di distribuzione** si definisce per variabili aleatorie discrete e continue. La usiamo quando vogliamo conoscere la **probabilità** che la variabile casuale assuma valori minori o uguali a x.

$$F_X(x) = p(X \leq x)$$

Per variabili discrete:

$$F(x) = \sum_{k \leq x} P(X = k)$$

Per variabili continue:

$$F(x) = \int_{-\infty}^{x} f(t)\, dt$$

Valore Atteso o Media

Il **valore atteso** (o **media**) di una variabile aleatoria rappresenta il valore medio che ci si aspetta di ottenere in un vasto numero di osservazioni dell'esperimento.

Se X variabile aleatoria discreta:

$$E[X] = \sum_i x_i P(X = x_i)$$

Se X variabile aleatoria continua:

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

Proprietà:

- **Linearità**: Se a e b sono costanti e X, Y variabili aleatorie:

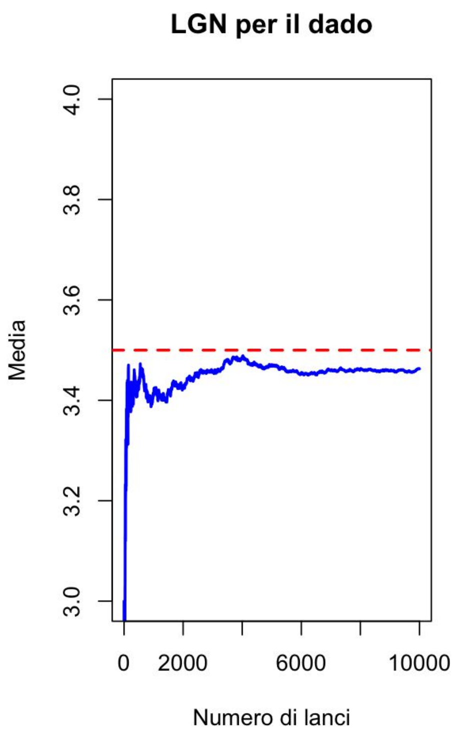
$$E[aX + bY] = aE[X] + bE[Y]$$

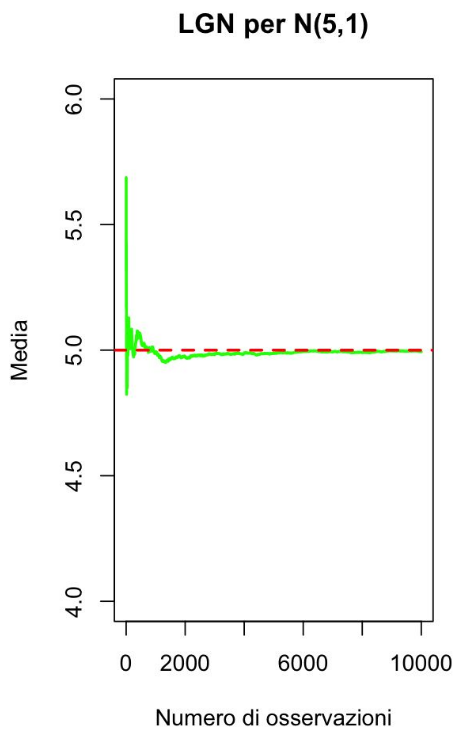
- Se X è simmetrica rispetto a 0, allora $E[X] = 0$.

Legge dei Grandi Numeri

Ripetendo un esperimento casuale un numero elevato di volte, la media aritmetica dei risultati osservati tende al valore atteso della variabile aleatoria.

$$\overline{X}_n = \frac{1}{n}\sum_{i=0}^{n} X_i \qquad \text{converge a } E[X] \text{ per } n \to \infty$$





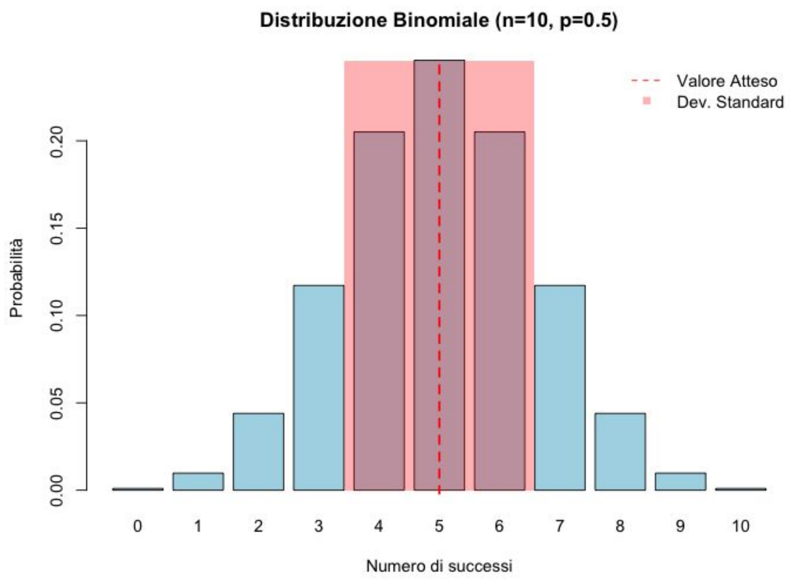
Varianza

La **varianza** di una variabile aleatoria misura la sua dispersione nei suoi valori attorno alla sua media $E[X]$.

$$Var(X) = E\left[(X - E[X])^2\right]$$

Deviazione standard:

$$\sigma = \sqrt{Var(X)}$$



V.A. Binomiali

$\mathfrak{B}(n, p)$

Si usa in caso di un esperimento con prove ripetute con le caratteristiche:

- Due soli esiti possibili: **successo** o **insuccesso**;
- Probabilità di successo (o insuccesso) costante;
- Risultati delle prove indipendenti.

Esempio:

- Lancio di una moneta;
- Tiro libero nel basket.

La variabile che descrive ogni prova è detta **binomiale** o di **Bernoulli**:

$$X = \begin{cases} 1 & P(X=1) = p \\ 0 & P(X=0) = 1-p \end{cases}$$

Dove $0 \leq p \leq 1$ è la probabilità del successo.

$$P(X=x) = \binom{n}{x} p^x (1-p)^{n-x}$$

X conta il numero di successi in n prove indipendenti ognuna con probabilità p.

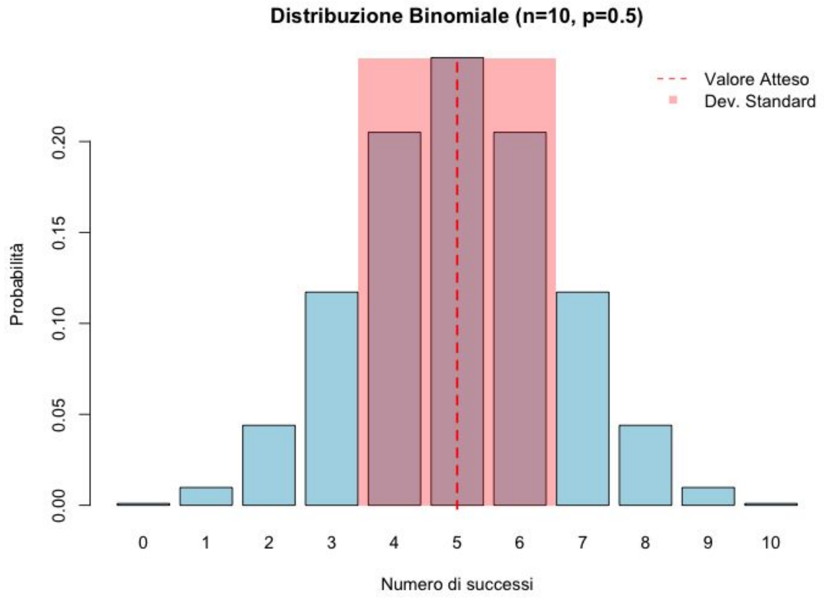
V.A. Binomiali

Valore Atteso o media:

$$E[X] = np$$

Varianza:

$$Var(X) = np(1 - p)$$



V.A. Geometriche

$\mathcal{G}(p)$

Si usa in caso di un esperimento con prove ripetute con le caratteristiche:

- Due soli esiti possibili: **successo** o **insuccesso**;
- Probabilità di successo (o insuccesso) costante;
- Risultati delle prove indipendenti;
- La variabile è il numero di prove necessarie ad ottenere il primo successo.

Esempio:

- Presa una catena di produzione, dopo quanti prodotti uno viene scartato perché difettoso?

La probabilità di avere un successo al k-esimo tentativo è data da:

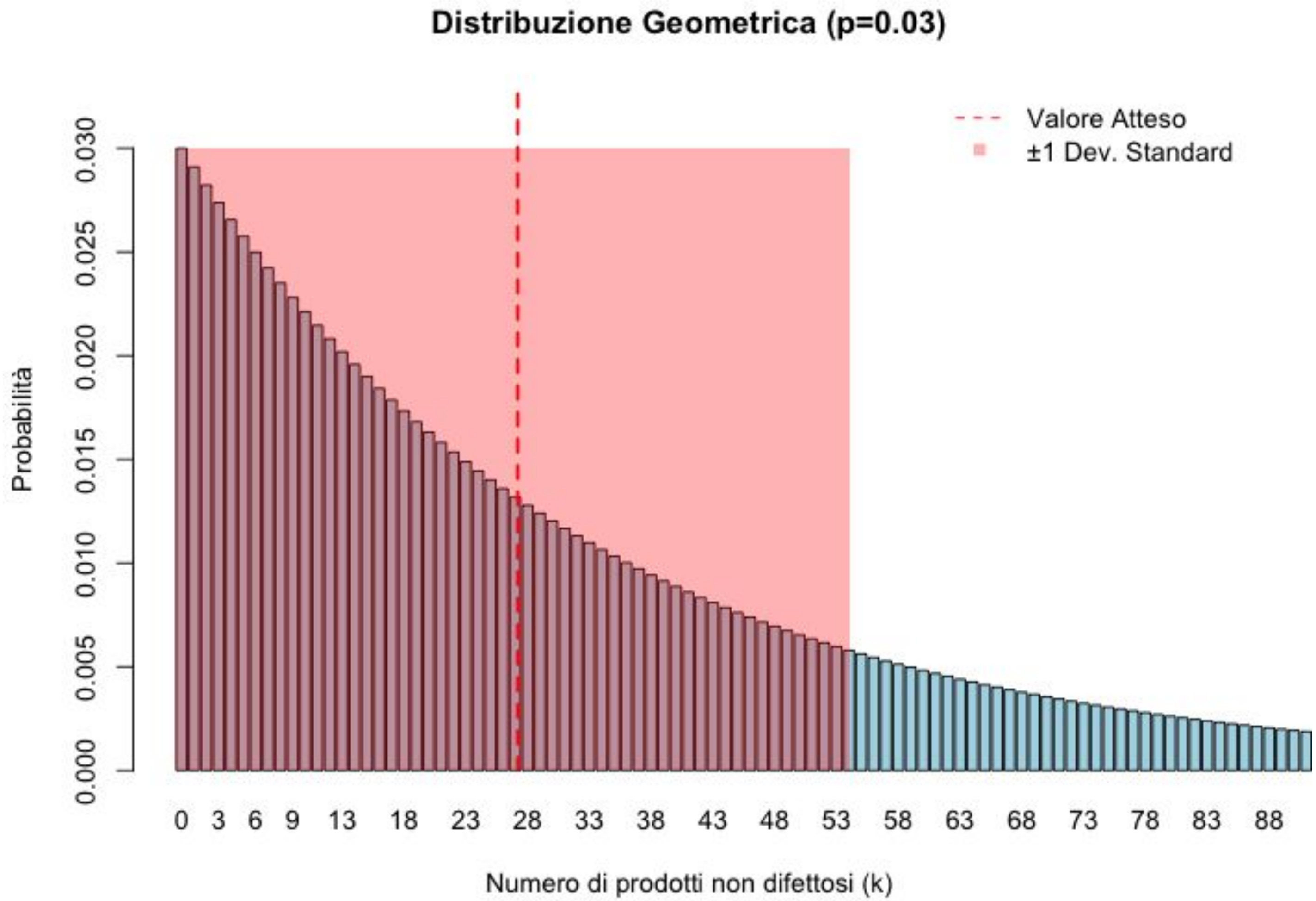
$$P(X=k) = p(1-p)^{k-1}$$

Valore Atteso o media:

$$E[X] = \frac{1}{p}$$

Varianza:

$$Var(X) = \frac{1-p}{p^2}$$


Distribuzione Geometrica (p=0.03)
Valore Atteso
±1 Dev. Standard
Probabilità
0.000
0.005
0.010
0.015
0.020
0.025
0.030
0 3 6 9 13 18 23 28 33 38 43 48 53 58 63 68 73 78 83 88
Numero di prodotti non difettosi (k)

V.A. di Poisson

$\mathscr{P}(\lambda)$

Si usa per ottenere la probabilità connessa al numero di eventi che si verificano in un determinato lasso di tempo, come

- Numero di incidenti in una giornata;
- Numero di chiamate ricevute da un centralino in 5 minuti.

Gli eventi si possono associare ad una variabile di **Poisson** se:

- Sono indipendenti tra di loro;
- La probabilità del verificarsi di un evento in un intervallo di tempo infinitesimo è proporzionale ad un parametro che caratterizza la prova;
- La probabilità che si verifichino più eventi nell'intervallo di riferimento è approssimabile a zero.

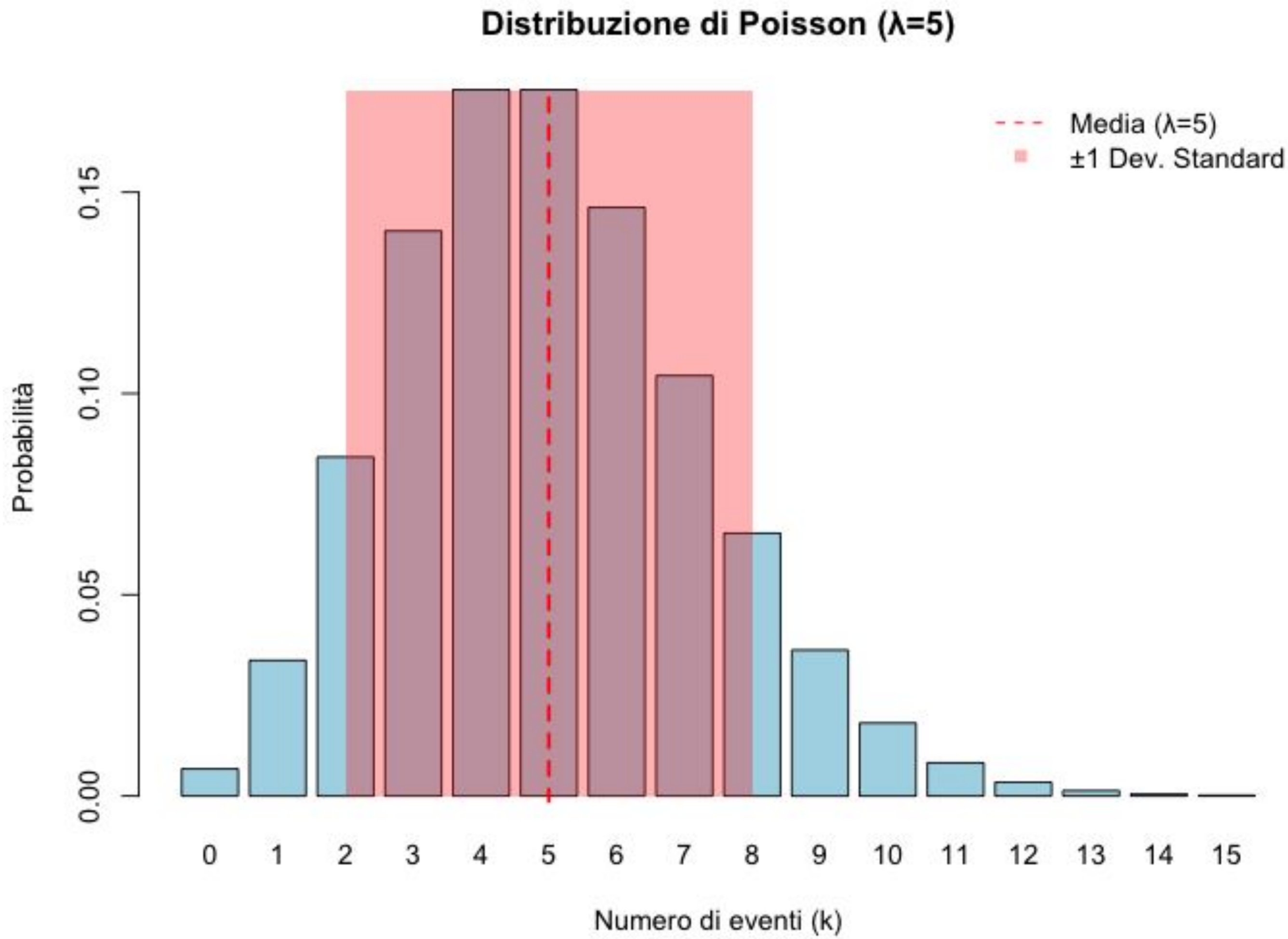
$$P(X=k)=\frac{\lambda^k \cdot e^{-\lambda}}{k!}$$

Valore Atteso o media:

$$E[X]=\lambda$$

Varianza:

$$Var(X)=\lambda$$


Distribuzione di Poisson (λ=5)
Media (λ=5)
±1 Dev. Standard
Probabilità
0.00
0.05
0.10
0.15
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
Numero di eventi (k)

V.A. Esponenziali

Le **variabili esponenziali** sono utilizzate per modellare il tempo tra eventi in un **processo di Poisson**, ovvero il tempo di attesa fino al verificarsi di un evento raro. Sono, per esempio, utilizzate in:

- **Teoria delle code**;
- **Affidabilità dei sistemi**.

$$f(x) = \lambda e^{-\lambda x}$$

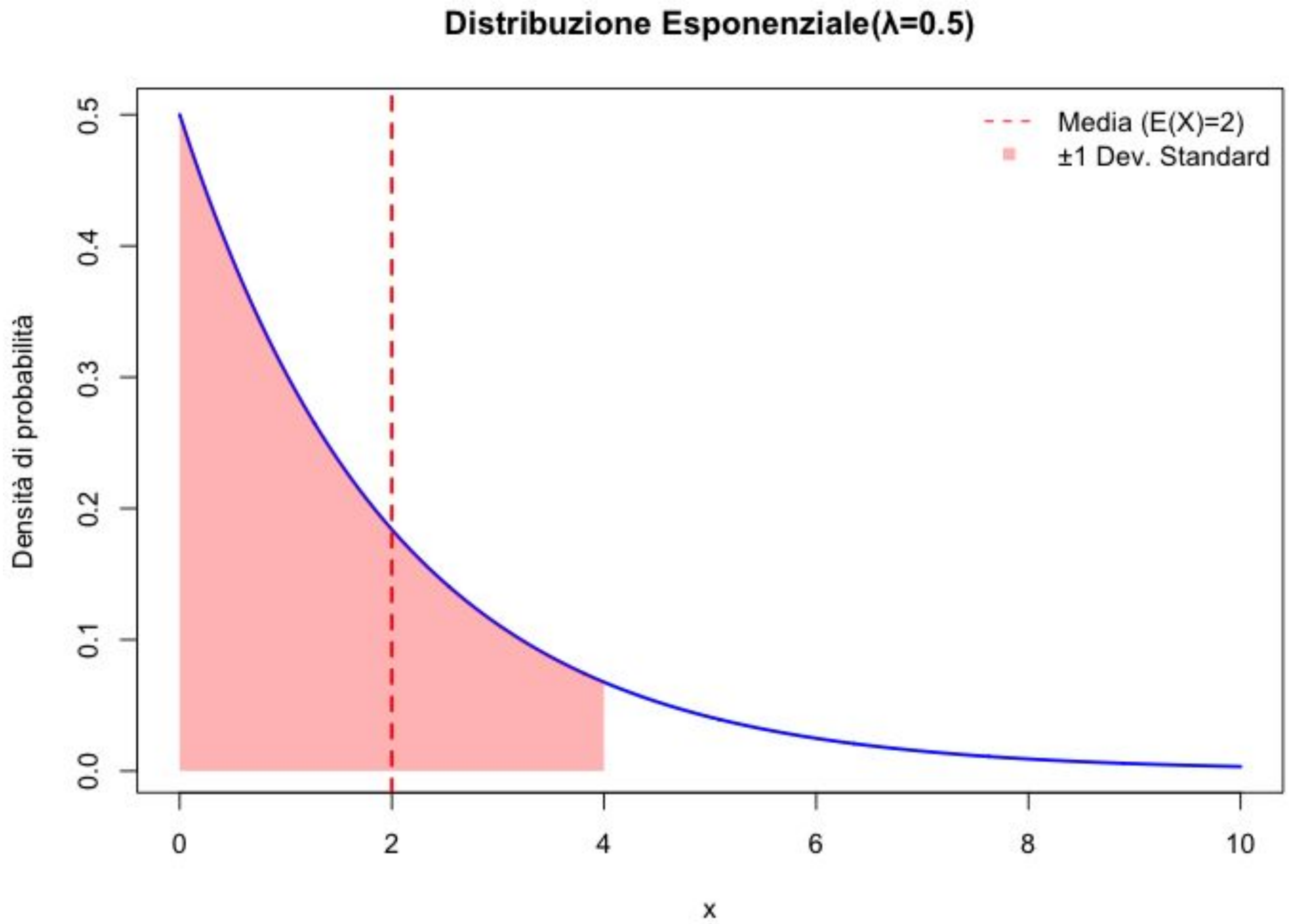
- $x \in (0, +\infty]$;
- λ numero di eventi per lasso di tempo.

Valore Atteso o media:

$$E[X] = \frac{1}{\lambda}$$

Varianza:

$$Var(X) = \frac{1}{\lambda^2}$$


Distribuzione Esponenziale(λ=0.5)
Media (E(X)=2)
±1 Dev. Standard
Densità di probabilità
0.0
0.1
0.2
0.3
0.4
0.5
0
2
4
6
8
10
x

Indipendenza e Correlazione

Due variabili X, Y si dicono **indipendenti** se, la conoscenza di una non fornisce informazioni rispetto all'altra.
Ovvero:

$$P(X=x, Y=y) = P(X=x)\,P(Y=y)$$

La **correlazione**, invece, misura la **dipendenza lineare**, tra due variabili, tramite il Coefficiente di Pearson:

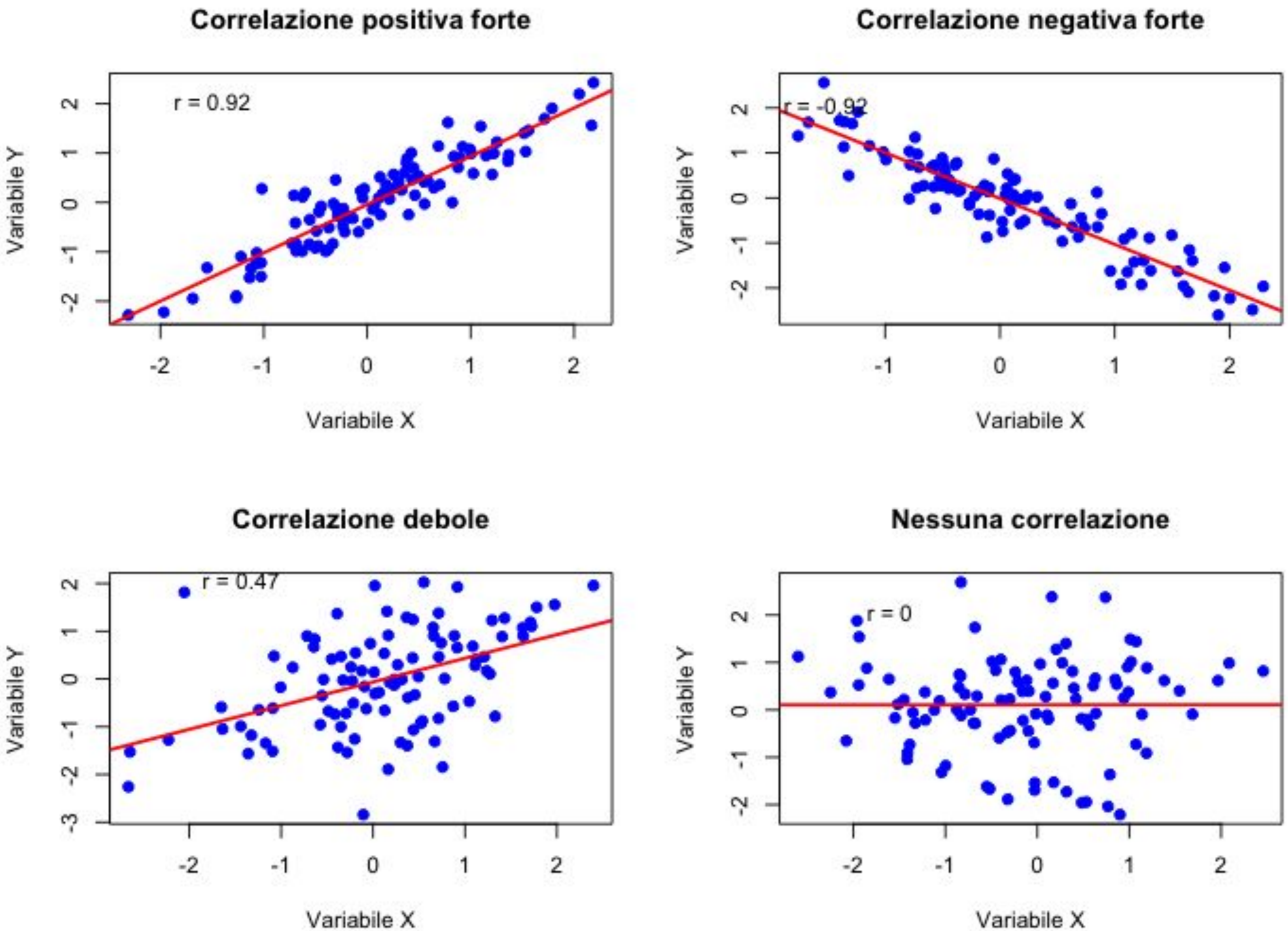
$$\rho(X,Y) = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$$

- $Cov(X,Y) = E[(X - E[X])(Y - E[Y])]$
- σ_X, σ_Y deviazione standard di X, Y

$$\rho(X,Y) \in [-1, 1]$$

$$Cov = 0 \Rightarrow \rho(X,Y) = 0$$

$$\rho(X,Y) = 0 \not\Rightarrow Cov = 0$$


Correlazione positiva forte
r = 0.92
Variabile Y
Variabile X
Correlazione negativa forte
r = -0.92
Variabile Y
Variabile X
Correlazione debole
r = 0.47
Variabile Y
Variabile X
Nessuna correlazione
r = 0
Variabile Y
Variabile X

V.A. Ipergeometriche

$\mathscr{H}(n, k, r)$

Viene utilizzata per stimare la probabilità di ottenere un certo numero di successi in un campione di dimensione fissa, estratto **senza reinserimento** da una popolazione finita.

$$P(X=k) = \frac{\binom{H}{k}\binom{N-H}{n-k}}{\binom{N}{n}}$$

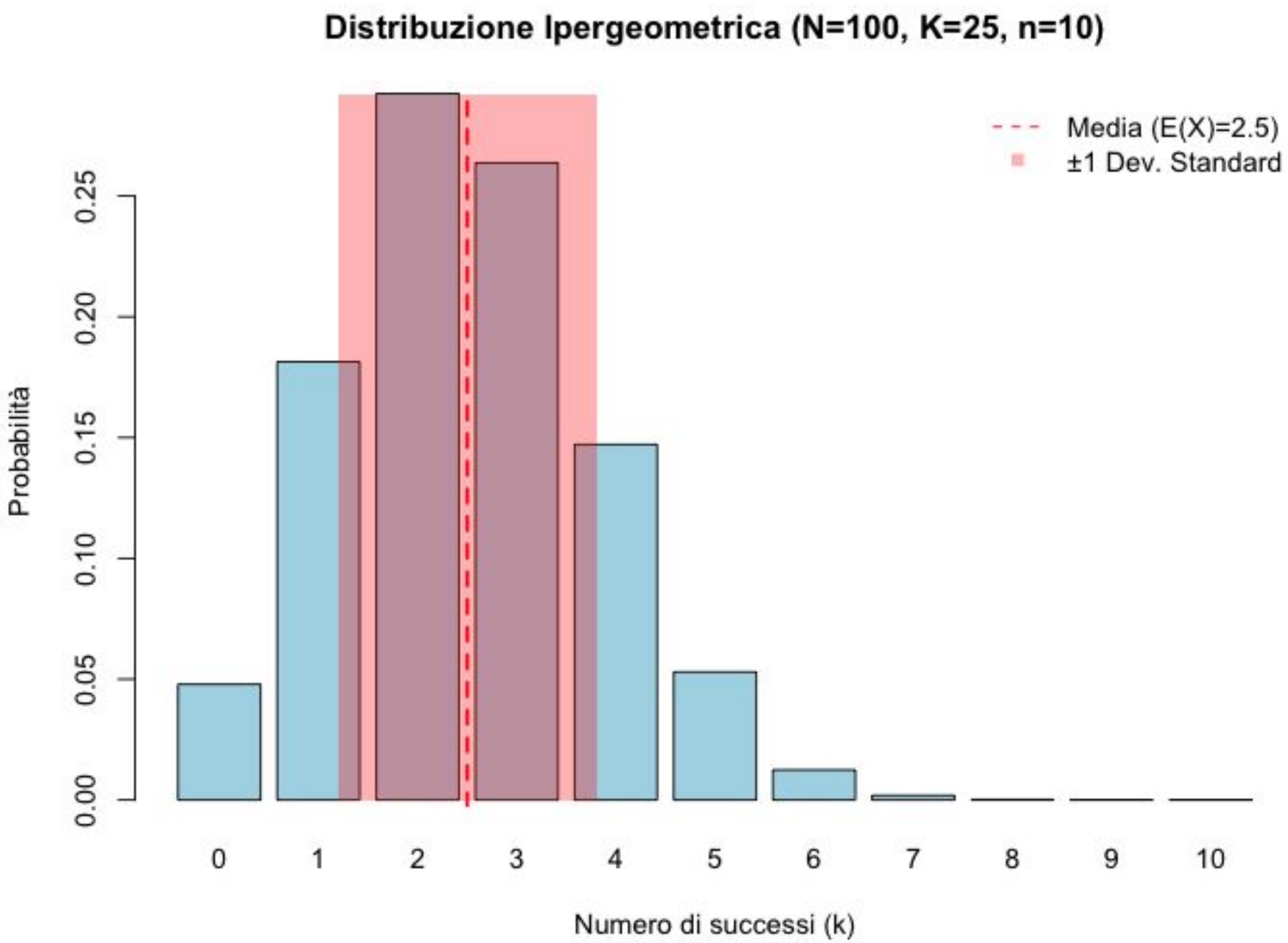
- H: numero di elementi di interesse;
- k: numero di successi in n prove;
- N: numero totale di elementi;
- n: numero di prove.

Valore Atteso o media:

$$E[X] = n\frac{H}{N}$$

Varianza:

$$Var(X) = np(1-p)\frac{N-n}{N-1}$$


Distribuzione Ipergeometrica (N=100, K=25, n=10)
Media (E(X)=2.5)
±1 Dev. Standard
Probabilità
0.00
0.05
0.10
0.15
0.20
0.25
0
1
2
3
4
5
6
7
8
9
10
Numero di successi (k)

V.A. Binomiali Negative o di Pascal

$\mathcal{N}\mathfrak{B}(p, n)$

Descrive il numero di fallimenti k ottenuti prima di un certo numero di successi r in una serie di prove indipendenti. Ciascuna con probabilità p di successo. Generalizza la distribuzione geometrica, dove $r = 1$.

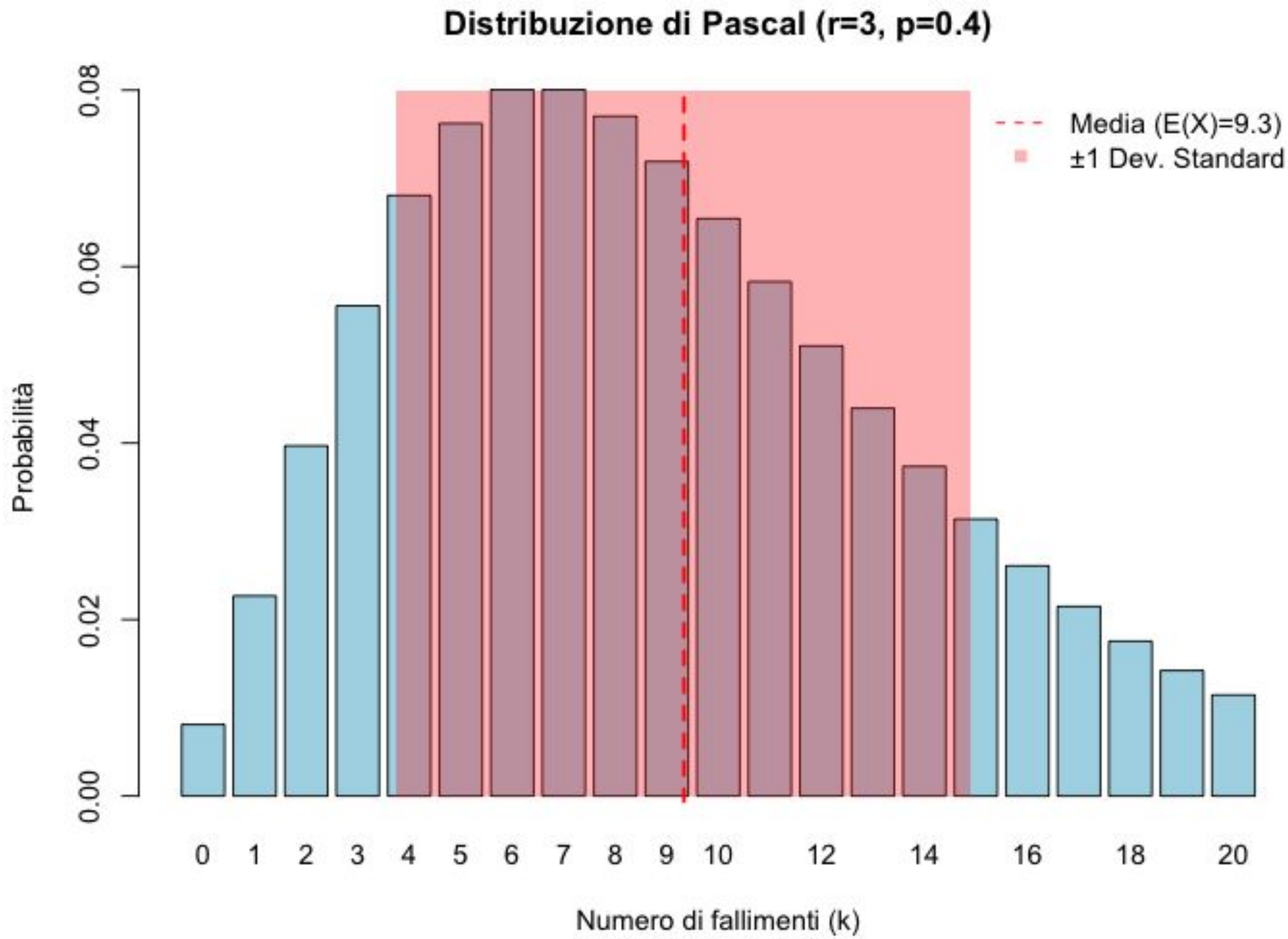
$$P(X = k) = \binom{k + r - 1}{k} p^r (1 - p)^k$$

Valore Atteso o media:

$$E[X] = \frac{r(1 - p)}{p}$$

Varianza:

$$Var(X) = \frac{r(1 - p)}{p^2}$$


Distribuzione di Pascal (r=3, p=0.4)
Media (E(X)=9.3)
±1 Dev. Standard
Probabilità
0.00
0.02
0.04
0.06
0.08
0 1 2 3 4 5 6 7 8 9 10 12 14 16 18 20
Numero di fallimenti (k)

V.A. Normali

$\mathcal{N}\left(\mu, \sigma^2\right)$

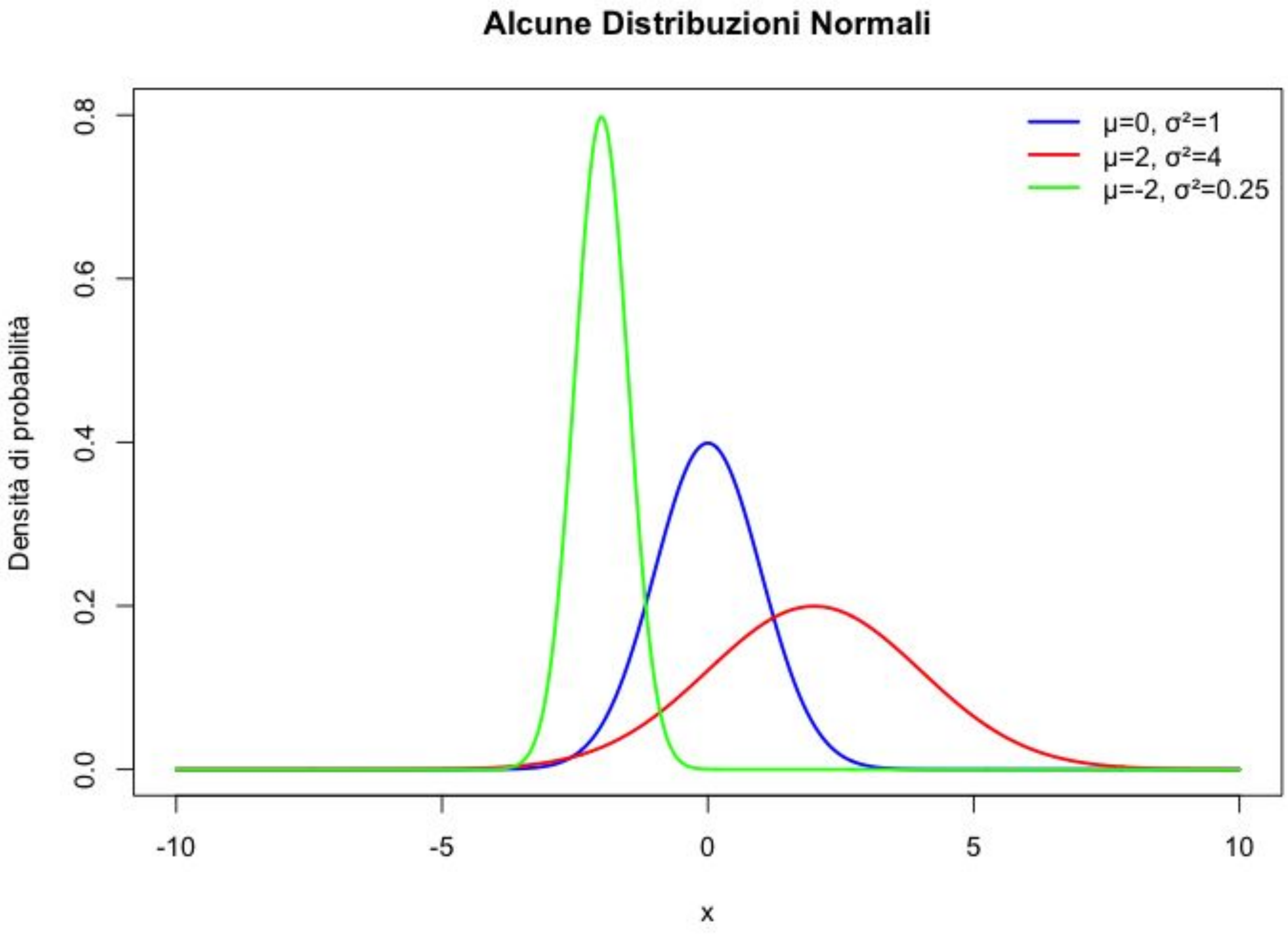
Le variabili normali sono una delle distribuzioni più utili in statistica e probabilità, viene usata per:

- Modellare fenomeni reali;
- Statistica inferenziale;
- Approssimazione di altre distribuzioni;
- **Standardizzazione** e confronto.

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Proprietà:

- **Forma a campana**;
- **Media**, **mediana**, **moda** coincidono;
- **Regola empirica**:
  - $\mu \pm \sigma$ copre circa il **68.26%** dei dati;
  - $\mu \pm 2\sigma$ copre circa il **95.45%** dei dati;
  - $\mu \pm 3\sigma$ copre circa il **99.73%** dei dati.


Alcune Distribuzioni Normali
μ=0, σ²=1
μ=2, σ²=4
μ=-2, σ²=0.25
Densità di probabilità
0.0
0.2
0.4
0.6
0.8
-10
-5
0
5
10
x

Normalizzazione di variabili

La **normalizzazione** è un processo che trasforma una variabile per renderla adatta all'analisi statistica o machine learning. Serve a:

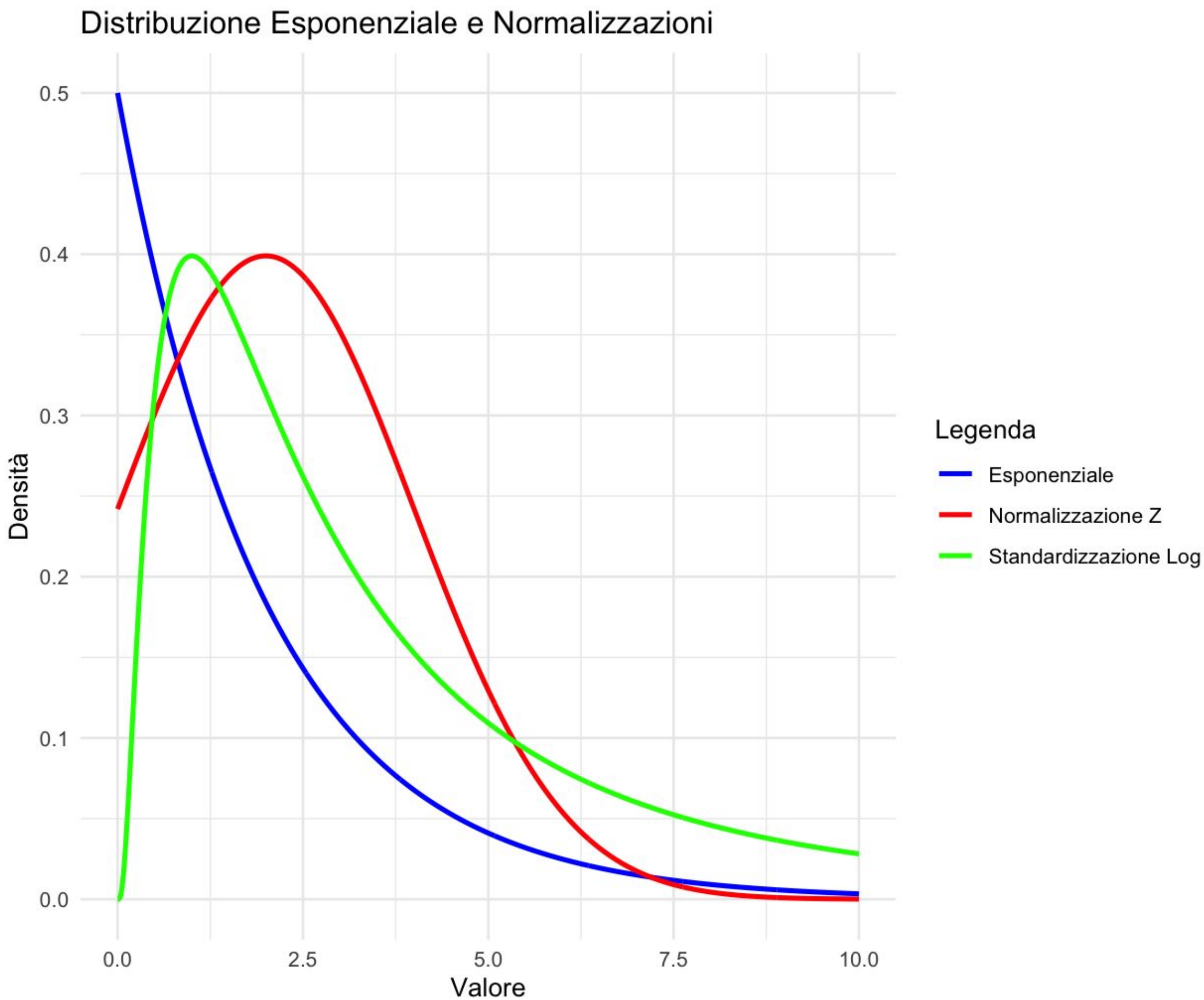
- **Rende variabili comparabili**;
- **Migliorare le prestazioni di algoritmi**;
- **Ridurre l'asimmetria** per renderla simile ad una normale.

Z-score:

$$z = \frac{x - \mu}{\sigma}$$

Trasformazione logaritmica:

$$z = \log(X)$$


Distribuzione Esponenziale e Normalizzazioni
Densità
Valore
0.0
0.1
0.2
0.3
0.4
0.5
0.0
2.5
5.0
7.5
10.0
Legenda
Esponenziale
Normalizzazione Z
Standardizzazione Log