Probabilità e Statistica[1]

Isadora Antoniano-Villalobos
isadora.antoniano@unive.it

Laurea in Informatica
(Data science/ Tecnologie e scienze dell'informazione)
Università Ca' Foscari di Venezia

Anno academico 2023/2024

[1] Materiale didattico redatto da: Isadora Antoniano-Villalobos & Federica Giummolè

Alcune variabili aleatorie continue

Distribuzione uniforme

Distribuzione uniforme

Immaginiamo che X sia una variabile che può assumere qualsiasi valore nell'intervallo (a, b), indifferentemente. La sua densità di probabilità è allora costante nell'intervallo (a, b) e nulla al di fuori:

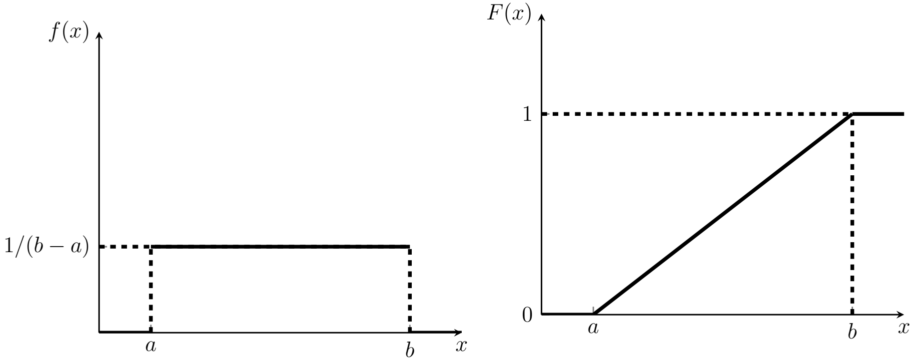
$$f(x) = \begin{cases} \frac{1}{b-a} & x \in (a,b) \\ 0 & \text{altrove} \end{cases}$$

X è una variabile aleatoria **uniforme** e si scrive $X \sim \mathrm{U}(a, b)$.

A partire dalla densità si può trovare la funzione di ripartizione:

$$F(x) = \int_a^x \frac{1}{b-a} dy = \frac{x-a}{b-a}; \quad x \in (a,b)$$

Distribuzione uniforme



E' facile verificare che

$$\mathbb{E}[X] = \frac{a+b}{2}, \qquad \mathrm{Var}[X] = \frac{(b-a)^2}{12}.$$

Distribuzione normale

Distribuzione normale

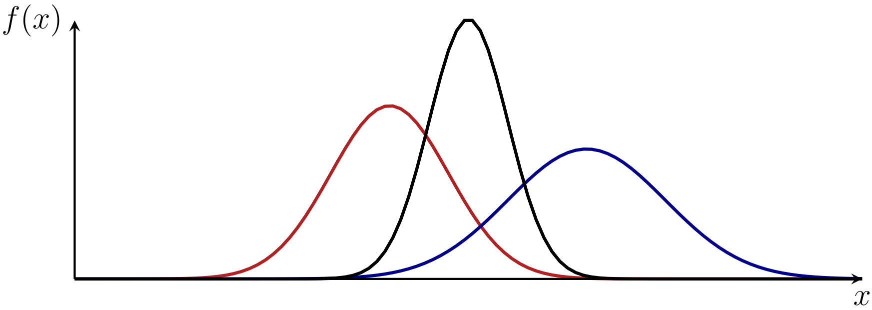
Una variabile aleatoria X ha **distribuzione normale** o gaussiana, indicato $X \sim \mathsf{N}(\mu, \sigma^2)$, se la sua funzione di densità ha la forma

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\}, \quad x \in \mathbb{R},\ \mu \in \mathbb{R},\ \sigma > 0.$$

☞ **Johann Friedrich Carl Gauss (1777–1855):** matematico, astronomo e fisico tedesco, che ha dato contributi determinanti in analisi matematica, teoria dei numeri, statistica, calcolo numerico, geometria differenziale, geodesia, geofisica, magnetismo, elettrostatica, astronomia e ottica. Talvolta definito *il più grande matematico della modernità*.(Wikipedia)

Carl Friedrich Gauss

Distribuzione normale

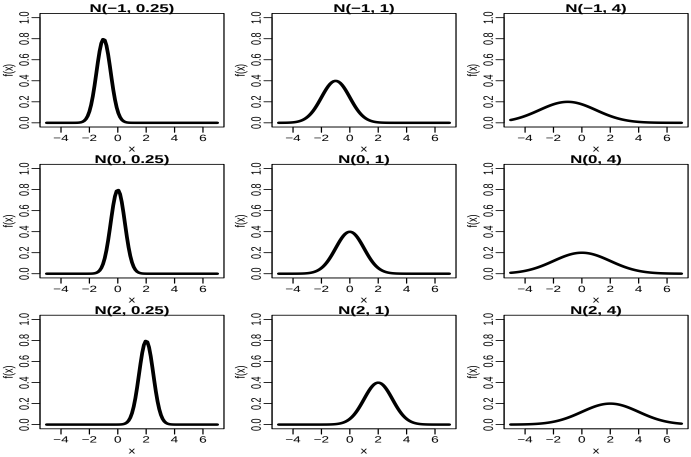


Si può dimostrare che

$$\mathbb{E}[X] = \mu, \qquad \text{Var}[X] = \sigma^2 \ .$$

Distribuzione normale

I parametri μ e σ^2 della normale non cambiano la forma a campana della densità ma determinano la sua posizione e concentrazione:



Distribuzione normale

La funzione di ripartizione di una v.a. normale non si può calcolare in forma esplicita. Ma utilizzando **R** (o un qualsiasi pacchetto statistico) si può oggi calcolare numericamente:

Esempio: $X \sim \mathrm{N}\left(-1, 4\right)$

➔ $\mathbb{P}\left[X \leq 1\right] = 0.8413$

```
> pnorm(q=1,mean=-1,sd=2)
[1] 0.8413447
```

➔ $\mathbb{P}\left[0 \leq X \leq 1\right] = 0.1499$

```
> pnorm(q=1,mean=-1,sd=2)-pnorm(q=0,mean=-1,sd=2)
[1] 0.1498823
```

Distribuzione normale

Standardizzazione

La variabile $Z \sim N(0,1)$ viene anche chiamata **normale standard** e la sua f.r. si denota con la Φ. Alla normale standard si riducono tutte le altre normali mediante la **standardizzazione**:

$$X \sim N(\mu, \sigma^2) \quad \Rightarrow \quad Z = \frac{X-\mu}{\sigma} \sim N(0,1)$$

Una volta (e ancora adesso a volte...), per calcolare delle probabilità per qualsiasi variabile aleatoria normale si usavano le **tavole della distribuzione normale standard** (Le trovate nel vostro libro: date un'occhiata!). Queste bastano grazie alla standardizzazione:

$$\mathbb{P}\left[a < X < b\right] = \mathbb{P}\left[\frac{a-\mu}{\sigma} < Z < \frac{b-\mu}{\sigma}\right] = \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right) .$$

Distribuzione normale

Esempio:

$$X \sim \mathsf{N}(-1,4) \quad \Rightarrow \quad Z = \frac{X+1}{2} \sim \mathsf{N}(0,1)$$

➔ $\mathbb{P}[X \leq 1] = \mathbb{P}\left[Z \leq \frac{1+1}{2}\right]$

```
> pnorm(q=1,mean=-1,sd=2)
[1] 0.8413447
> pnorm(q=(1+1)/2,mean=0,sd=1)
[1] 0.8413447
```

➔ $\mathbb{P}[0 \leq X \leq 1] = \mathbb{P}\left[\frac{0+1}{2} \leq Z \leq \frac{1+1}{2}\right]$

```
> pnorm(q=1,mean=-1,sd=2)-pnorm(q=0,mean=-1,sd=2)
[1] 0.1498823
> pnorm(q=(1+1)/2,mean=0,sd=1)-pnorm(q=(0+1)/2,mean=0,sd=1)
[1] 0.1498823
```

Distribuzione normale

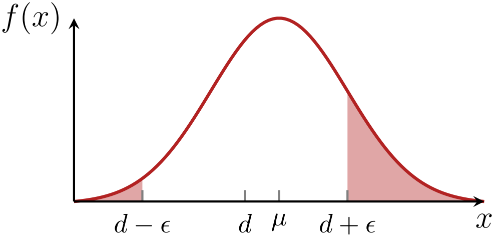
Esempio: Una macchina produce tubi di diametro X (in mm). Il diametro X potrebbe essere modellizzato con una distribuzione normale di media μ mm e varianza σ^2 mm^2. Supponiamo che per contratto i tubi debbano essere di diametro d mm più o meno un margine di specifica di ϵ mm. Si noti che d non è necessariamente uguale a μ perché la macchina potrebbe essere centrata su valori leggermente diversi.

- Calcolare la probabilità che un generico tubo sia difettoso (cioè fuori specifica, o non conforme).
  - ☞ Sia $p = \mathbb{P}\left[\text{“X è difettoso”}\right]$. Allora,

$$p = \mathbb{P}\left[(X < d - \epsilon) \cup (d + \epsilon < X)\right] = \mathbb{P}\left[X < d - \epsilon\right] + \mathbb{P}\left[d + \epsilon < X\right]$$
$$= \int_{-\infty}^{d-\epsilon} \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\} dx + \int_{d+\epsilon}^{+\infty} \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\} dx.$$

Distribuzione normale

Dati μ, σ, T e ϵ siamo in grado di calcolare p (in R, per esempio, con il comando `pnorm`).



- Calcolare la probabilità che due tubi su 10 siano difettosi.
  - Sia
  $$S_{10} = \text{“numero di tubi difettosi su 10”}$$
  la variabile che conta il numero di successi in 10 prove indipendenti in cui ogni prova può risultare un successo (difettoso o non conforme) con probabilità
  $$p = \mathbb{P}\left[\text{“X è difettoso”}\right]$$

Distribuzione normale

o un insuccesso (conforme) con probabilità $1 - p$. Allora

$$S_{10} \sim \mathsf{Bin}\,(10, p)$$

e la probabilità che due tubi su 10 siano difettosi è data dalla formula della funzione di probabilità binomiale:

$$\mathbb{P}\left[S_{10} = 2\right] = \binom{10}{2} p^2 (1-p)^8.$$

Esercizio: Fare i calcoli con $\mu = 3$, $\sigma = 0.09$, $d = 2.94$ e $\epsilon = 0.18$.

☞ $p = 0.0950$, $\mathbb{P}\left[S_{10} = 2\right] = 0.1827$

Approssimazione normale per la binomiale

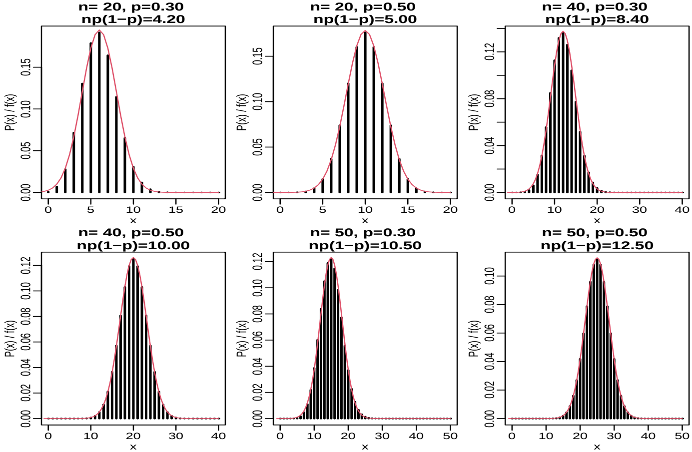
Quando n è "grande", allora la funzione di ripartizione di una v.a. binomiale di parametri n e p si può approssimare con la funzione di ripartizione di una normale di parametri $\mu = np$ e $\sigma^2 = np(1-p)$:

Approssimazione normale per la binomiale

$$\mathsf{Bin}\,(n, p) \approx \mathsf{N}\,(np, np(1-p))\,.$$

- L'approssimazione viene utilizzata nella pratica quando $np(1-p) \geq 10$.
- Questa è una conseguenza del **teorema del limite centrale** (che studieremo più avanti)

Approssimazione normale per la binomiale



Approssimazione normale per la binomiale

Esempio: Un certo virus danneggia un file con probabilità 0.35, indipendentemente dagli altri file. Il virus attacca una cartella con 2400 file. Qual è la probabilità che vengano danneggiati fra gli 800 e gli 850 file (inclusi)?

- Sia X la variabile che conta il numero di file danneggiati su 2400. Allora $X \sim \text{Bin}\,(2400, 0.35)$ e la probabilità richiesta è:

$$\mathbb{P}\left[800 \leq X \leq 850\right] = \sum_{k=800}^{850} \binom{2400}{k} 0.35^k (1-0.35)^{2400-k} = 0.632893.$$

Calcolata con **R**:

```
> pbinom(850,2400,0.35)-pbinom(799,2400,0.35)
[1] 0.632893
```

Approssimazione normale per la binomiale

Possiamo anche approssimare X con una normale con parametri

$$\mu = 2400 \cdot 0.35 = 840$$

$$\sigma = \sqrt{2400 \cdot 0.35 \cdot (1 - 0.35)} = \sqrt{546} = 23.36664.$$

Otteniamo, con la cosiddetta **correzione per continuità**:

$$\begin{aligned}
\mathbb{P}\left[800 \leq X \leq 850\right] &= \mathbb{P}\left[\mathbf{799.5} < X < \mathbf{850.5}\right] \\
&\approx \Phi\left(\frac{850.5 - 840}{23.36664}\right) - \Phi\left(\frac{799.5 - 840}{23.36664}\right) \\
&= \Phi(0.4493586) - \Phi(-1.73324) = 0.631887.
\end{aligned}$$

Calcolata con **R**:

```
>  pnorm((850.5-840)/23.36664)-pnorm((799.5-840)/23.36664)
[1] 0.631887
```

Distribuzione gamma

Distribuzione gamma

Si dice che X ha **distribuzione gamma** di parametri $\alpha > 0$ e $\lambda > 0$, e si scrive $X \sim \text{Ga}\,(\alpha, \lambda)$, se la sua densità di probabilità è della forma

$$f(x) = \begin{cases} \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x} & x \in (0, \infty) \\ 0 & \text{altrove} \end{cases}$$

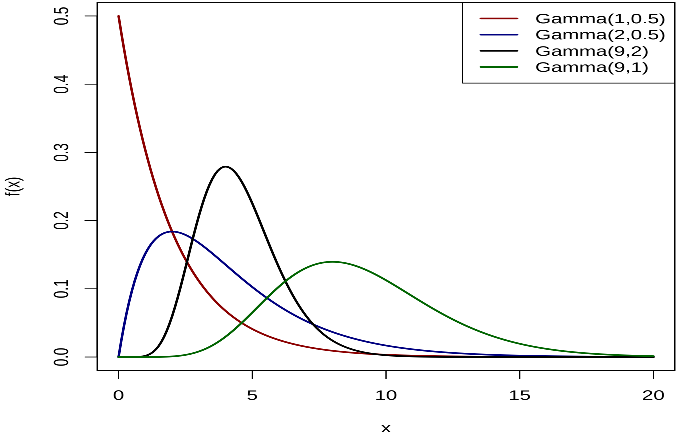
dove $\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x}\, dx$ è la **funzione gamma**.

Si può dimostrare che

$$\mathbb{E}\left[X\right] = \frac{\alpha}{\lambda}, \qquad \text{Var}\left[X\right] = \frac{\alpha}{\lambda^2}\ .$$

Distribuzione gamma

La densità della gamma al variare di α e λ cambia sia forma che posizione:



Distribuzione gamma

Esempio: Se X è una v.a. gamma di media 4 e varianza 4, allora

$$\alpha = 4 \quad \text{e} \quad \lambda = 1 \ .$$

- Si può allora calcolare (integrando ripetutamente per parti)

$$\begin{aligned} \mathbb{P}\left[X < 4\right] &= \int_0^4 \frac{e^{-x}x^3}{6}\, dx \\ &= \frac{1}{6}\left[-(x^3 + 3x^2 + 6x + 6)e^{-x}\right]_0^4 = \\ &= 0.567 \ . \end{aligned}$$

- In questo caso il calcolo esplicito è stato possibile poiché α è un numero intero e

$$\Gamma(4) = \int_0^\infty x^{4-1}e^{-x}\, dx = (4-1)! = 6 \ .$$

Distribuzione gamma

In generale, usando **R** si possono calcolare probabilità legate a variabili con distribuzione gamma:

- Per $X \sim \text{Ga}\,(1.5, 0.5)\,, \quad \mathbb{P}\,[X \leq 1] = 0.19874804$

```
> pgamma(q=1,shape=1.5,rate=0.5)
[1] 0.198748
```

- Per $X \sim \text{Ga}\,(7.5, 1)\,, \quad \mathbb{P}\,[X \geq 7.2] = 0.49543391$

```
> 1-pgamma(q=7.2,shape=7.5,rate=1)
[1] 0.4954339
```

Si può anche valutare la funzione gamma:

- $\Gamma(4) = 6 \quad \Gamma(7.5) = 1871.254$

```
> gamma(4)
[1] 6
> gamma(7.5)
[1] 1871.254
```

Distribuzione esponenziale

Distribuzione esponenziale

Un caso particolare di distribuzione gamma con $\alpha = 1$ è la **distribuzione esponenziale**, $X \sim \text{Exp}\,(\lambda)$.

La sua densità è

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & x \in (0, \infty) \\ 0 & \text{altrove} \end{cases}$$

e la sua funzione di ripartizione si può calcolare in forma esplicita:

$$F(x) = \begin{cases} 0 & x < 0 \\ 1 - e^{-\lambda x} & x \geq 0 \end{cases}$$

Distribuzione esponenziale

Evidentemente,

$$\mathbb{E}\left[X\right] = \frac{1}{\lambda} \quad \text{e} \quad \mathsf{Var}\left[X\right] = \frac{1}{\lambda^2}\ .$$

La distribuzione esponenziale si usa per modellare tempi di attesa. Ad esempio:

1. il tempo che passa fra l'arrivo di un treno in stazione e il successivo;
2. la vita di un certo componente elettronico di un'automobile;
3. la lunghezza fra due difetti consecutivi di una fibra ottica.

Distribuzione esponenziale

Mancanza di memoria

Se $X \sim \mathrm{Exp}(\lambda)$, allora

$$\begin{aligned}\mathbb{P}[X > t + s | X > t] &= \frac{\mathbb{P}[X > t + s]}{\mathbb{P}[X > t]} \\ &= \mathbb{P}[X > s] \ .\end{aligned}$$

Questa proprietà si dimostra tenendo conto che

$$\mathbb{P}[X > x] = e^{-\lambda x} \ , \quad x > 0 \ .$$

- La distribuzione esponenziale è l'unica distribuzione continua con la proprietà di **mancanza di memoria** (come la geometrica fra le distribuzioni discrete).

Distribuzione esponenziale

Esempio: Il tecnico di un laboratorio dell'università in un'ora installa un certo *software* in media su 30 pc. Assumendo che il tempo di installazione su ogni pc segua una distribuzione esponenziale, vogliamo calcolare la probabilità che il tecnico impieghi più di 5 minuti per installare il *software* nel prossimo pc.

- X=tempo di attesa in ore $\sim Exp(30)$

$$\Rightarrow \mathbb{P}\left[X > 1/12\right] = e^{-30/12}$$

- X=tempo di installazione in minuti $\sim Exp(1/2)$

$$\Rightarrow \mathbb{P}\left[X > 5\right] = e^{-5/2}$$

Il processo di Poisson

Ricordiamo: La variabile di Poisson viene spesso utilizzata per i conteggi.

- **Esempio:** i messaggi in arrivo nella mia casella di posta elettronica sono in media 12 ogni 10 minuti. Inoltre si può supporre che la variabile che conta i messaggi in arrivo in un intervallo di 10 minuti abbia distribuzione di Poisson:

$$X_{10} = \text{“numero messaggi in 10 min.”} \sim \mathcal{P}(12)$$

Possiamo allora calcolare probabilità del tipo:

$$\mathbb{P}\left[\text{meno di 6 messaggi fra le 11:00 e le 11:10}\right] = \\ = \mathbb{P}\left[X_{10} < 6\right] = 0.02\ .$$

- E se fossimo interessati a intervalli di tempo di durata diversa da 10 minuti?

Il processo di Poisson

Processo di Poisson

Il processo di Poisson è una successione di variabili aleatorie $\{X_t\}_{t \geq 0}$, con distribuzione di Poisson il cui parametro dipende dall'indice t:

$$X_t \sim \mathcal{P}(\lambda \times t).$$

Si usa per contare il numero di manifestazioni di un fenomeno di interesse in un qualsiasi intervallo di tempo di ampiezza t, quando il numero medio di queste manifestazioni dipende soltanto dall'ampiezza (e non dalla posizione) dell'intervallo considerato.

Possiamo allora scrivere:

- $X_t =$"n. di eventi in un intervallo di tempo t"
- $\lambda =$"n. medio di eventi nell'unità di tempo".

Il processo di Poisson

Esempio: Il numero di messaggi in arrivo alla mia casella di posta elettronica segue un processo di Poisson nel tempo, con una media di 12 messaggi ogni 10 minuti.

☞ Se scegliamo il minuto come unità di misura del tempo, abbiamo:

$$\lambda = \text{"n. medio di messaggi in 1 min."} = 12/10$$

$$X_t = \text{"n. di messaggi in } t \text{ min."} \sim \text{Po}\left(\frac{12}{10}t\right)$$

Possiamo allora calcolare, ad esempio,

$$\mathbb{P}\left[\text{più di 20 messaggi in mezz'ora}\right] = \\ = \mathbb{P}\left[X_{30} > 20\right] = 1 - \mathbb{P}\left[X_{30} \leq 20\right],$$

con $X_{30} \sim \text{Po}\left(\frac{12}{10}30\right) = \text{Po}\,(36)$.

Il processo di Poisson

Relazione con la distribuzione esponenziale: Associata ad ogni processo di Poisson c'è una variabile aleatoria esponenziale che misura il **tempo fra due manifestazioni successive** del fenomeno in questione:

- $X_t =$"n. di eventi in un intervallo di tempo t"

$$X_t \sim \mathsf{Po}\,(\lambda\, t)$$

- $T =$"tempo trascorso fra due eventi successivi"

$$T \sim \mathsf{Exp}\,(\lambda)$$

Il processo di Poisson

Esempio: Per l'esempio dei messaggi in arrivo alla casella di posta elettronica:

$$
\begin{aligned}
&\mathbb{P}\left[\text{nessun messaggio in 15 minuti}\right] = \\
&= \mathbb{P}\left[X_{15} = 0\right] = \frac{(12/10 \times 15)^0}{0!} e^{-12/10 \times 15} = e^{-18} \\
&= \mathbb{P}\left[T > 15\right] = e^{-12/10 \times 15} = e^{-18},
\end{aligned}
$$

dove $T \sim \text{Exp}\left(12/10\right)$ misura il tempo che passa fra l'arrivo di due messaggi successivi.