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

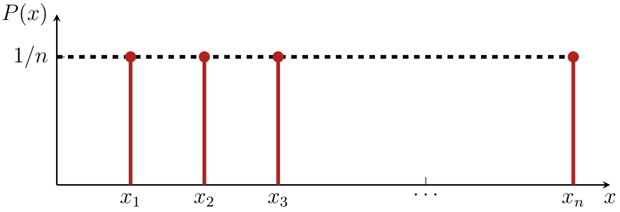
Alcune variabili aleatorie discrete

Distribuzione Uniforme Discreta

Distribuzione Uniforme (Discreta)

Consideriamo una variabile aleatoria X che assume un numero finito di valori, $\{x_1, \ldots, x_n\}$, tutti con la stessa probabilità $p_i = 1/n$, $i = 1, \ldots, n$. Si dice allora che X ha una **distribuzione uniforme** e si scrive

$$X \sim \mathsf{U}\{x_1, \ldots, x_n\}$$



Distribuzione Uniforme Discreta

Esempi:

1. X descrive il risultato del lancio di un dado (non truccato). I valori possibili per X sono gli interi fra 1 e 6, ciascuno con probabilità $1/6$ di verificarsi.

2. Si fa una scommessa in cui si guadagnano 1000 euro se lanciando una moneta equilibrata si ottiene testa e se ne perdono 1500 se il risultato è croce. La variabile X che indica il guadagno ottenuto, ha distribuzione uniforme sull'insieme $\{x_1 = 1000, x_2 = -1500\}$.

Distribuzione ipergeometrica

Riconsideriamo il campionamento da una popolazione divisa in due sottopopolazioni. Questo schema costituisce un modello applicabile a molte situazioni differenti:

- difettoso/non difettoso in controllo della qualità;
- sotto soglia/sopra soglia in rilevazioni di inquinamento;
- destra/sinistra in *polls* elettorali;
- positivo/negativo in un test.

Conoscendo la composizione della popolazione (proporzione di individui con la proprietà di interesse) e il tipo di campionamento (con o senza reinserimento), abbiamo imparato a calcolare la probabilità di eventi del tipo:

$$A_k = \text{“}k \text{ elementi su } n \text{ hanno la proprietà richiesta”}$$

.

Distribuzione ipergeometrica

Spesso si indica come **successo** l'estrazione di un individuo con la proprietà richiesta e l'evento A_k diventa "k successi su n estrazioni". Possiamo ora interpretare questi risultati utilizzando le variabili aleatorie.

Distribuzione ipergeometrica

Sia X la variabile aleatoria che conta il numero di successi su n estrazioni senza reinserimento da una popolazione con N elementi dei quali K sono considerati successo. Si dice allora che X ha una **distribuzione ipergeometrica** di parametri N, K e n e si scrive

$$X \sim \mathsf{Ig}\,(N, K, n)$$

Si ha che

$$\mathbb{P}\left[X = k\right] = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}\ ,$$

per $k = \max\{0, n-(N-K)\}, \ldots, \min\{n, K\}$, con $n \leq N$.

Distribuzione ipergeometrica

Si può dimostrare (ma noi non lo faremo) che:

- La media della distribuzione ipergeometrica è:

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

- La varianza della distribuzione ipergeometrica è:

$$\mathsf{Var}[X] = n\frac{K}{N}\frac{N-K}{N}\frac{N-n}{N-1}\ .$$

Distribuzione ipergeometrica

➔ **Esempio:** Un *software* consiste di 12 programmi, 5 dei quali necessitano di *upgrade*. Se vengono scelti a caso 4 programmi per un test

1. qual è la probabilità che almeno 2 di essi siano da aggiornare?
   Sia X la variabile che conta il numero di programmi da aggiornare fra i 4 scelti.
   Allora $X \sim \text{Ig}\,(N = 12, K = 5, n = 4)$ e la probabilità richiesta è

$$\begin{aligned} \mathbb{P}\left[X \geq 2\right] &= 1 - \mathbb{P}\left[X \leq 1\right] \\ &= 1 - \mathbb{P}\left[X = 0\right] - \mathbb{P}\left[X = 1\right] \\ &= 1 - \frac{\binom{5}{0}\binom{12-5}{4-0}}{\binom{12}{4}} - \frac{\binom{5}{1}\binom{12-5}{4-1}}{\binom{12}{4}} \;. \end{aligned}$$

2. qual è il numero medio di programmi da aggiornare fra i 4 scelti?
   Il numero atteso di programmi da aggiornare fra i 4 scelti è

$$\mathbb{E}\left[X\right] = 4\frac{5}{12} = \frac{5}{3} \;.$$

Distribuzione di Bernoulli

Supponiamo di eseguire un esperimento i cui possibili risultati possono essere classificati come *successo*, con probabilità p o *insuccesso*, con probabilità $1-p$. Questo tipo di esperimento viene comunemente chiamato **prova bernoulliana**.

- **Jakob Bernoulli (1654–1705):** Noto anche come Jacques, James o Giacomo Bernoulli, è stato un matematico e scienziato svizzero. La sua opera principale, *Ars Conjectandi*, pubblicata postuma nel 1713, è un lavoro fondamentale per la teoria delle probabilità. I concetti *campionamento bernoulliano*, *teorema di Bernoulli*, *variabile casuale bernoulliana* e *numeri di Bernoulli* sono legati ai suoi lavori e nominati in suo onore. Inoltre il primo teorema centrale del limite, ovvero la legge dei grandi numeri, venne formulata da Jakob. (Wikipedia)

Jakob Bernoulli

Distribuzione di Bernoulli

Distribuzione di Bernoulli

Sia X la variabile casuale che prende il valore $X = 1$ quando l'esito di una prova bernoulliana è un successo e $X = 0$ quando questo è un insuccesso. Si dice allora che X ha una **distribuzione di Bernoulli** di parametro $p \in (0, 1)$ e si scrive

$$X \sim \text{Ber}\,(p)$$

Si ha che,

$$\mathbb{P}\left[X = x\right] = p^x(1-p)^{1-x}\mathbf{1}_{\{0,1\}}(x) = \begin{cases} 1-p & \text{se } x = 0; \\ p & \text{se } x = 1; \\ 0 & \text{altrimenti} \end{cases}$$

$$\mathbb{E}\left[X\right] = p, \quad \text{Var}\left[X\right] = p(1-p)$$

Distribuzione di Bernoulli

Esempio: L'estrazione di un singolo individuo da una popolazione con N elementi dei quali K sono considerati successi è una prova bernoulliana. La variabile casuale $X =$ "Numero di successi" ha una distribuzione di Bernoulli con parametro $p = K/N$.

☞ Anche la variabile casuale $Y =$ "Numero di insuccessi" ha una distribuzione di Bernoulli, ma con parametro $q = 1 - p = (N - K)/N$.

Distribuzione binomiale

Distribuzione binomiale

Supponiamo di eseguire n prove bernoulliane indipendenti, ognuna con probabilità di successo p. Sia X la variabile aleatoria che conta il numero totale di successi ottenuti nelle n prove. Si dice allora che X ha una **distribuzione binomiale** di parametri n e $p \in (0,1)$ e si scrive

$$X \sim \mathsf{Bin}\,(n, p)$$

Si ha che

$$\mathbb{P}\,[X = x] = \binom{n}{x} p^x (1-p)^{n-x}, \quad x = 0, 1, \ldots, n\ .$$

Distribuzione binomiale

Si osservi che si tratta effettivamente di una distribuzione di probabilità:

$$\begin{aligned}\sum_{x=0}^{n} \mathbb{P}\left[X = x\right] &= \sum_{x=0}^{n} \binom{n}{x} p^x (1-p)^{n-x} \\ &= (p + 1 - p)^n = 1,\end{aligned}$$

per la formula del binomio di Newton.

- Sia X la variabile aleatoria che conta il numero di successi su n estrazioni con reinserimento da una popolazione con N elementi dei quali K sono considerati successo. La probabilità di successo rimane invariata, uguale a $p = K/N$, ad ogni estrazione successiva. Quindi, X ha una distribuzione binomiale di parametri n e $p = K/N$.
- La distribuzione di Bernoulli è un caso particolare della binomiale con $n = 1$.

Distribuzione binomiale

Esempio: Si consideri la solita urna con 4 palline bianche e 3 nere. Sia S_3 la variabile che conta il numero di palline bianche ottenute in tre estrazioni con reinserimento. Qual è la $\mathbb{P}[S_3 = 2]$?

- ➔ Considerando l'estrazione di pallina bianca come successo e visto che $\mathbb{P}[\text{bianca}] = 4/7$, si può affermare che $S_3 \sim \text{Bin}(3, 4/7)$.
  Allora,
  $$\mathbb{P}[S_3 = 2] = \binom{3}{2} \left(\frac{4}{7}\right)^2 \left(\frac{3}{7}\right)^{(3-2)} = 0.4198 \; .$$

La distribuzione binomiale si utilizza anche nel caso di **Campionamento da popolazioni infinite** $(N \to \infty)$, quando si conosce ad ogni estrazione la probabilità di successo p.

- ☞ Se la popolazione viene considerata infinita, non si distingue nemmeno fra estrazioni con o senza reinserimento perché si assume che in ogni caso la probabilità di successo in estrazioni successive non cambia, e si usa sempre la binomiale.

Distribuzione binomiale

Esempio: Un motore di ricerca su internet cerca una parola in una sequenza (finita, ma così numerosa da considerarsi infinita) di siti web indipendenti. Si pensa che il 20% di questi siti possa contenere la parola cercata. Qual è la probabilità che 5 dei primi 10 siti visitati contengano la parola cercata? E la probabilità che al meno 5 dei primi 10 siti la contengano?

- Sia Y la variabile che conta il numero di siti, fra i primi 10 visitati, che contengono la parola cercata. Allora $Y \sim \text{Bin}\,(n = 10, p = 0.2)$ e

$$\mathbb{P}\,[Y = 5] = \binom{10}{5} 0.2^5 (1 - 0.2)^{10-5} = 0.0264$$

$$\begin{aligned}\mathbb{P}\,[Y \leq 5] &= \mathbb{P}\,[Y = 0] + \mathbb{P}\,[Y = 1] + \ldots + \mathbb{P}\,[Y = 5] \\ &= 0.1074 + 0.2684 + \ldots + 0.0264 = 0.9936\end{aligned}$$

Distribuzione binomiale

Usando R:

➔ $\mathbb{P}\left[Y=0\right]$

```
> dbinom(x=0,size=10,prob=0.2)
[1] 0.1073742
```

➔ $\mathbb{P}\left[Y=1\right]$

```
> dbinom(x=1,size=10,prob=0.2)
[1] 0.2684355
```

$\vdots$

➔ $\mathbb{P}\left[Y=5\right]$

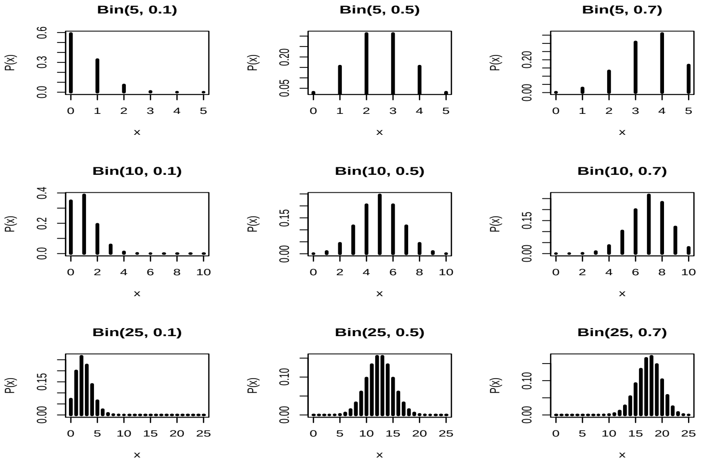
```
> dbinom(x=5,size=10,prob=0.2)
[1] 0.02642412
```

➔ O direttamente, $\mathbb{P}\left[Y\leq 5\right]$

```
> pbinom(q=5,size=10,prob=0.2)
[1] 0.9936306
```

Distribuzione binomiale

I parametri della binomiale determinano la forma della distribuzione:



Distribuzione binomiale

- Il valore atteso di $X \sim \text{Bin}\,(n, p)$ è

$$\begin{aligned}
\mathbb{E}\left[X\right] &= \sum_{k=0}^{n} k \binom{n}{k} p^k (1-p)^{n-k} \\
&= \sum_{k=1}^{n} k \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k} \\
&= np \sum_{k=1}^{n} \frac{(n-1)!}{(k-1)!(n-k)!} p^{k-1} (1-p)^{n-k} \\
&= np \sum_{k=0}^{n-1} \frac{(n-1)!}{k!(n-1-k)!} p^k (1-p)^{n-1-k} = np\ .
\end{aligned}$$

Distribuzione binomiale

- La varianza di X si calcola tenendo conto che

$$\begin{aligned}\mathbb{E}\left[X^2\right] &= \sum_{k=0}^{n} k^2 \binom{n}{k} p^k (1-p)^{n-k} \\ &= \sum_{k=0}^{n} (k^2 - k) \binom{n}{k} p^k (1-p)^{n-k} + \mathbb{E}\left[X\right] \\ &= n(n-1)p^2 + np \ .\end{aligned}$$

Si ha dunque

$$\text{Var}\left[X\right] = \mathbb{E}\left[X^2\right] - \left[\mathbb{E}\left[X\right]\right]^2 = np(1-p) \ .$$

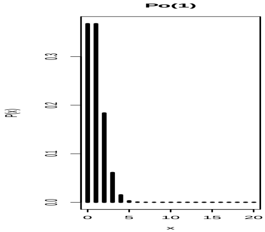
Distribuzione di Poisson

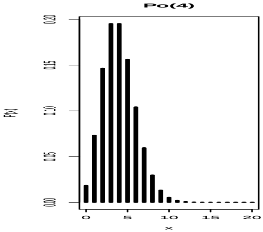
Distribuzione di Poisson

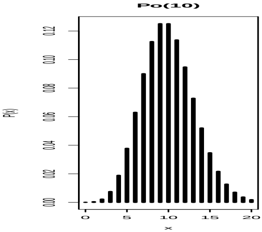
Una variabile X che assume valori nell'insieme dei numeri naturali $\mathbb{N}$, ha **distribuzione di Poisson** di parametro $\lambda > 0$ se

$$\mathbb{P}\left[X = k\right] = \frac{\lambda^k}{k!}\,e^{-\lambda}, \quad k = 0, 1, 2, \ldots \ .$$

Scriveremo allora $X \sim \mathrm{Po}\left(\lambda\right)$







Distribuzione di Poisson

☞ E' facile vedere che $\sum_i p_i = 1$:

$$\sum_{i=0}^{+\infty} \frac{\lambda^i}{i!} e^{-\lambda} = e^{-\lambda} \sum_{i=0}^{+\infty} \frac{\lambda^i}{i!} = e^{-\lambda} e^{\lambda} = 1 \ .$$

La variabile di Poisson viene utilizzata come modello per il conteggio di manifestazioni di un certo fenomeno di interesse. Ad esempio:

1. chiamate in arrivo ad un centralino in un certo intervallo di tempo;
2. macchine transitanti ad un casello autostradale in un certo periodo del giorno;
3. difetti rilevati in un pezzo di filo d'acciaio prodotto da una ditta;
4. terremoti manifestatisi in una data area nell'arco degli ultimi 10 anni.

Distribuzione di Poisson

- Il valore atteso di $X \sim \mathcal{P}(\lambda)$ è

$$\begin{aligned}\mathbb{E}[X] &= \sum_{i=0}^{+\infty} i \frac{\lambda^i}{i!} e^{-\lambda} \\ &= \lambda \sum_{i=1}^{+\infty} \frac{\lambda^{i-1}}{(i-1)!} e^{-\lambda} = \lambda \ .\end{aligned}$$

- Allo stesso modo si dimostra che

$$\mathsf{Var}[X] = \lambda \ .$$

Distribuzione di Poisson

Esempio: Al mio account di posta elettronica arrivano messaggi con una media di 10 ogni mezz'ora. Si può inoltre supporre che il numero di messaggi in arrivo segua una distribuzione di Poisson. Qual è la probabilità che nella prossima mezz'ora mi arrivino non più di 3 messaggi? E almeno 12?

➔ $X =$"n. messaggi ogni mezz'ora"$\sim \text{Po}\,(\lambda = 10)$.

$$\begin{aligned}\mathbb{P}\left[X \leq 3\right] =& \mathbb{P}\left[X = 0\right] + \mathbb{P}\left[X = 1\right] + \mathbb{P}\left[X = 2\right] + \mathbb{P}\left[X = 3\right] \\ =& \frac{10^0}{0!}\,e^{-10} + \frac{10^1}{1!}e^{-10} + \frac{10^2}{2!}\,e^{-10} + \frac{10^3}{3!}\,e^{-10} \\ =& 0.0103\end{aligned}$$

➔ **Esercizio:** $\mathbb{P}\left[X \geq 12\right] = 1 - \mathbb{P}\left[X \leq 11\right] = \ldots \quad .$

Distribuzione di Poisson

Calcolo delle probabilità:

- Utilizzando **R** (o un qualsiasi pacchetto statistico) si può oggi lavorare con le principali funzioni di probabilità, di densità e di ripartizione.
- Una volta (e ancora adesso a volte...) si usavano invece delle tabelle preconfezionate, le **tavole della distribuzione**. Nelle tavole sono registrati i valori di alcune funzioni di ripartizione relativi a diversi valori dei parametri caratterizzanti la distribuzione. Alcune tavole sono nell'appendice del vostro libro: date un'occhiata!

Esempio:

$$X \sim \text{Po}(10)$$

➔ $\mathbb{P}[X = 12] = 0.0948$

```
> dpois(x=12, lambda = 10)
```

Distribuzione di Poisson

```
[1] 0.09478033
```

➔ $\mathbb{P}[X \leq 12] = 0.7916$

```
> ppois(q=12, lambda = 10)
[1] 0.7915565
```

➔ $\mathbb{P}[X > 12] = 1 - 0.7916 = 0.2084$

```
> 1-ppois(q=12, lambda = 10)
[1] 0.2084435
```

➔ $\mathbb{P}[X \geq 12] = \mathbb{P}[X > 11] = 1 - 0.6968 = 0.3032$

```
> 1-ppois(q=11, lambda = 10)
[1] 0.3032239
```

Approssimazione Poisson per la binomiale

Quando $n \to \infty$ e $p \to 0$ ma in modo tale che il prodotto $np \to \lambda$ rimane costante, allora la funzione di probabilità di una v.a. binomiale di parametri n e p si può approssimare con la funzione di probabilità di una Poisson di parametro λ:

Approssimazione Poisson per la binomiale

$$\lim_{\substack{n\to\infty \\ p\to 0 \\ np\to\lambda}} \binom{n}{x} p^x (1-p)^{n-x} = e^{-\lambda}\frac{\lambda^x}{x!}$$

- L'approssimazione viene utilizzata nella pratica quando $n \geq 100$ e $p \leq 0.05$. Si sceglie allora $\lambda = \mathbb{E}[X] = np$.

Approssimazione Poisson per la binomiale

- Questa è nota anche come la **legge degli eventi rari**, perchè si può interpretare come la distribuzione del numero di eventi di tipo "successo" quando gli esperimenti bernoulliani sono tanti $(n \to \infty)$ ma la probabilità di successo è molto piccola $(p \to 0)$, ovvero, i successi sono rari.

Esempio: Una fabbrica di componenti elettronici fornisce il 3% dei chip acquistati da un produttore di telefoni cellulari. Qual è la probabilità che su 100 chip acquistati ve ne siano al massimo 3 provenienti da quella fabbrica?

$X =$ Numero di chip provenienti dalla fabbrica su 100. Then $X \sim \text{Bin}\,(100, 0.03)$ e:

$$\mathbb{P}\left[X \leq 3\right] = \sum_{k=0}^{3} \binom{100}{k} 0.03^k 0.97^{100-k} = 0.64724921$$

☞ È difficile da calcolare senza computer!

Approssimazione Poisson per la binomiale

Dato che $n = 100$ è grande e $p = 0.03$ è piccolo, si può utilizzare l'approssimazione con la Poisson di parametro $\lambda = np = 3$,

$$\sum_{k=0}^{3} \frac{3^k}{k!} e^{-3} = 0.64723189,$$

che è più facile da calcolare e molto precisa.

Distribuzione geometrica

Distribuzione geometrica

Sia X una variabile aleatoria che conta il numero di ripetizioni indipendenti necessarie per osservare il primo successo in un esperimento binario che ha probabilità di successo p. Si dice allora che X ha una **distribuzione geometrica** di parametro $p \in (0,1)$ e si scrive

$$X \sim \mathsf{Geo}\,(p)$$

Si ha che

$$\mathbb{P}\left[X = x\right] = (1-p)^{x-1}p, \quad x = 1, 2, 3, \dots \ .$$

Si tratta effettivamente di una distribuzione di probabilità:

$$\sum_{x=1}^{\infty} \mathbb{P}\left[X = x\right] = p \sum_{x=1}^{\infty} (1-p)^{x-1} = p\,\frac{1}{1-(1-p)} = 1.$$

Distribuzione geometrica

Esercizio: Si dimostri che

- $\mathbb{E}[X] = 1/p$
- $\text{Var}[X] = (1-p)/p^2$.

Mancanza di memoria

Se $X \sim \text{Geo}(()\,p)$, allora

$$\mathbb{P}[X > m+n | X > m] = \frac{\mathbb{P}[X > m+n]}{\mathbb{P}[X > m]}$$
$$= \mathbb{P}[X > n] \ .$$

Questa proprietà si dimostra tenendo conto che

$$\mathbb{P}[X > k] = (1-p)^k \ , \quad k = 0, 1, 2, \ldots \ .$$

Distribuzione geometrica

☞ La distribuzione geometrica è l'unica distribuzione discreta con la proprietà di *mancanza di memoria*.

Esempio Si consideri ancora il motore di ricerca su internet. Se il 20% dei siti visitati contengono la parola cercata,

1. qual è la probabilità di dover visitare 15 siti per trovare la parola?
2. Dato che i primi 4 siti visitati non contenevano la parola cercata, qual è la probabilità di doverne visitare più di 10 in tutto per trovare la parola cercata?
3. Qual è il numero medio di siti da visitare per trovare la parola la prima volta?

Distribuzione geometrica

- Supponiamo che X conti il numero di pagine da visitare per trovare per la prima volta la parola cercata.
  Allora $X \sim \text{Geo}\,(0.2)$ e

$$\mathbb{P}\,[X = 15] = 0.2 \times 0.8^{14} = 0.0088,$$

$$\mathbb{P}\,[X > 10 | X > 4] = \mathbb{P}\,[X > 10 - 4 = 6] = 0.8^{6} = 0.2621,$$

$$\mathbb{E}\,[X] = \frac{1}{0.2} = 5.$$

Distribuzione geometrica

Attenzione! In alcuni contesti, si definisce la variabile aleatoria geometrica come il numero di ripetizioni indipendenti necessarie **prima di** osservare il primo successo. In questo caso, i possibili valori della variabile sono $\{0, 1, 2, \ldots\}$. Questa è la definizione usata da **R**. Infatti:

➔ Ricordiamo che $\mathbb{P}[X = 15] = 0.008796093$,

```
> dgeom(x=15, prob=0.2)
[1] 0.007036874
> dgeom(x=15-1, prob=0.2)
[1] 0.008796093
```