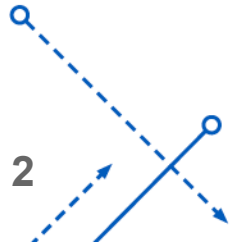
STATISTICA E ANALISI DEI DATI

Distribuzione Poisson

Distribuzione di Poisson

- La distribuzione di Poisson si utilizza è una distribuzione casuale discreta di **media e varianza identiche** che si utilizza per calcolare:
  - la probabilità che un certo evento si manifesti esattamente x volte in una **certa unità spazio-temporale**
- La distribuzione di Poisson interviene spesso nella descrizione di alcuni fenomeni coinvolgenti qualche tipo di conteggio
  - Il numero di chiamate telefoniche ricevute da un centralino in un fissato intervallo di tempo
  - Il numero di particelle radioattive emesse per unità di tempo
  - Il numero di microorganismi per unità di volume in un fluido
  - Il numero di imperfezioni per unità di lunghezza di un cavo
- Tutti questi eventi devono essere riferiti ad una certa unità temporale (e/o spaziale) e in questa conosciamo il numero medio di volte (che chiameremo lambda) che si manifesta un certo evento



Distribuzione di Poisson

- L'unità spazio-temporale (λ) è quindi un elemento essenziale per poter applicare la distribuzione di Poisson
- Se prendiamo a riferimento un certo evento e il numero medio di volte in cui questo si manifesta **dobbiamo conoscere con esattezza** anche il tempo o eventualmente lo **spazio** in cui questo si manifesta
- Quando ad esempio diciamo:
  - In un call center chiamano 6 persone all'ora
  - In un negozio entrano 10 persone in una mattinata (4 ore)
  - Il giocatore segna 20 gol all'anno (24 partite)

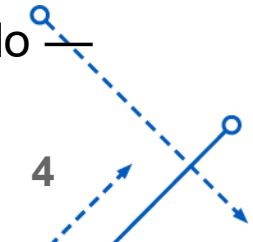
Distribuzione di Poisson

- Sia X la variabile aleatoria con:

Funzione di probabilità: $p_X(x) = P(X = x) = \begin{cases} \frac{\lambda^x}{x!} e^{-\lambda} & \lambda > 0 \\ 0 & altrimenti \end{cases}$

con $e = 2{,}718281 \ldots$ (numero di Nepero) si dice avere distribuzione Poisson di parametro λ

- λ^x: rappresenta il contributo proporzionale alla probabilità di osservare x eventi, ottenuto moltiplicando insieme x probabilità elementari
  - Esempio: Se in media arrivano $\lambda = 3$ chiamate al minuto, $\lambda^2 = 3^2 = 9$ rappresenta il "contributo proporzionale" del verificarsi di 2 chiamate
- $x!$ tiene conto del **numero di modi diversi in cui gli eventi possono essere ordinati**
  - Esempio: Se $x = 3$ eventi si verificano, potremmo osservarli in 3!=6 modi distinti (esempio: [A, B, C], [A, C, B], [B, A, C], [B, C, A], [C, A, B], [C, B, A])
  - Poiché la distribuzione di Poisson considera solo *quanti* eventi avvengono in un intervallo — e non *in quale ordine*, dobbiamo dividere per $x!$ per evitare di contare più volte la stessa configurazione di 3 eventi



Distribuzione di Poisson

- Sia X la variabile aleatoria con:

$$\textbf{Funzione di probabilità: } p_X(x) = P(X = x) = \begin{cases} \frac{\lambda^x}{x!} e^{-\lambda} & \lambda > 0 \\ 0 & altrimenti \end{cases}$$

con $e = 2{,}718281 \ldots$ (numero di Nepero) $0 < p < 1$ si dice avere distribuzione Poisson di parametro λ

  - Consideriamo $e^{-\lambda}$ rappresenta la probabilità che non avvenga alcun evento nell'intervallo considerato
    - Dal punto di vista matematico, questo deriva dal limite $(1 - \frac{\lambda}{n})^n \to e^{-\lambda}$,che rappresenta la probabilità che tutti i piccoli sotto-intervalli *non* contengano un evento
    - Nella formula della Poisson, $e^{-\lambda}$ funge da fattore di base e garantisce la normalizzazione della distribuzione
    - Combinato con $\lambda^x/x!$,produce la decrescita delle probabilità per valori di x molto lontani dal valore medio λ
    - Il termine $e^{-\lambda}$ riduce la probabilità di eventi più numerosi rispetto al valore medio λ, riflettendo che eventi molto lontani dal valore atteso (λ) sono sempre meno probabili

- Notazione: $X \sim P(\lambda)$ indicherà che X è una variabile aleatoria avente distribuzione di Poisson di parametro λ

Distribuzione di Poisson

- Notazione: $X \sim P(\lambda)$ indicherà che X è una variabile aleatoria avente distribuzione di Poisson di parametro λ
- Dalla funzione di probabilità si ricava:

$$\frac{p_X(x)}{p_X(x-1)} = \frac{\lambda}{x} \qquad x = 1,2, \dots$$

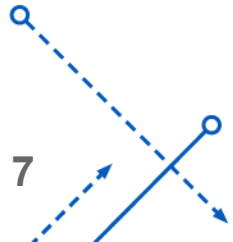
così che le probabilità di Poisson sono calcolabili in modo ricorsivo:

- La probabilità per x è: $P(X = x) = \frac{\lambda^x}{x!} e^{-\lambda}$
- La probabilità per $x - 1$ è: $P(X = x - 1) = \frac{\lambda^{x-1} e^{-\lambda}}{(x-1)!}$
- Dividendo $P(X = x)$ per $P(X = x - 1)$ si ha: $\frac{P(X=x)}{P(X=x-1)} = \frac{\frac{\lambda^x}{x!}}{\frac{\lambda^{x-1}}{(x-1)!}} = \frac{\lambda}{x}$
- Isolando $P(X = x)$ si ha: $P(X = x) = \frac{\lambda}{x} P(X = x - 1)$

- $P(X = 0) = e^{-\lambda}$
- $P(X = 1) = \frac{\lambda}{1} P(X = 0)$
- $P(X = 2) = \frac{\lambda}{2} P(X = 1)$
- …
- $P(X = k) = \frac{\lambda}{k} P(X = k - 1)$

Distribuzione di Poisson

- Il **numero di volte il cui l'evento si verifica** nell'unità spazio temporale è la media di questa distribuzione
  - Ritornando all'esempio di prima:
    - in un call center chiamano 6 persone all'ora
  - Intendiamo dire:
    - in un call center chiamano mediamente 6 persone in un'ora
- Il valore di λ e il valore dell'unità spazio-temporale sono tra di loro strettamente connessi
  - Non ha senso parlare del λ senza la corrispondente unità spazio-temporale
    - in un call center chiamano mediamente 6 persone in **un'ora**
  - Se **raddoppiamo il tempo** in cui l'evento si manifesta mediamente, andremo a **raddoppiare anche il numero di persone**
  - Viceversa, se dividiamo il tempo per 2, divideremo anche il numero di persone medio per 2



Esempio Giocatore

- Esempio: sappiamo che un giocatore è in grado di segnare mediamente 20 gol nell'arco di una stagione (24 partite)

  Qual è la probabilità che nelle prossime 6 partite giocate segni esattamente 6 goal?

  - $\lambda = 20$
  - Tempo = 24 partite

- Ora dobbiamo notare che la domanda ci chiede di calcolare la probabilità che il giocatore segni esattamente 6 gol nell'arco delle **prossime 6 partite** giocate
- Adeguiamo il numero medio di gol a 6 partite, ovvero un quarto di stagione:
  - $\lambda = \frac{20}{4} = 5$ con $\frac{1}{4} = \frac{6}{24}$ cioè $\frac{1}{4}$ di stagione
  - Tempo = $\frac{24}{4}$ = 6 partite
- Si ha che $p_X(x) = \frac{\lambda^x}{x!} e^{-\lambda} = \frac{5^6}{6!} e^{-5} = 0{,}1462$ cioè il 14,62% di probabilità che segni 6 goal nelle prossime 6 partite

Esempio Giocatore

- Per ricavare i valori della distribuzione di probabilità applichiamo sempre la formula:

$$p_X(x) = \frac{\lambda^x}{x!} e^{-\lambda}$$

e consideriamo x come il "numero di gol segnati da Ibrahimovic", si ha che con $\lambda = 5$:

$$p_X(0) = \frac{5^0}{0!} e^{-5}$$

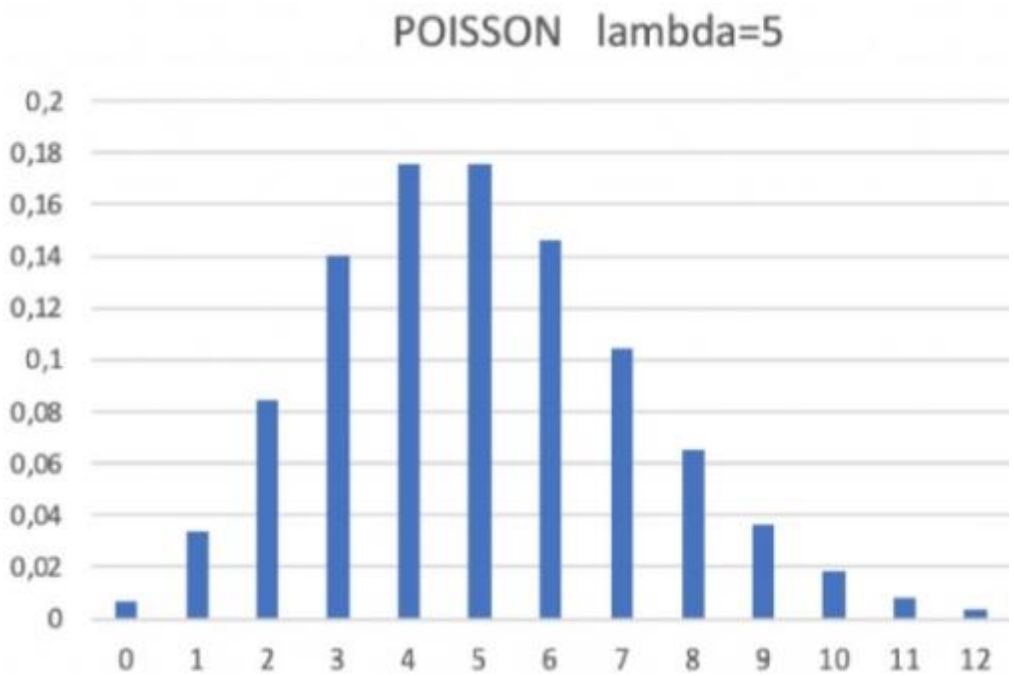
…

$$p_X(2) = \frac{5^2}{2!} e^{-5}$$

…

$$p_X(10) = \frac{5^{10}}{10!} e^{-5}$$

…



Distribuzione di Poisson

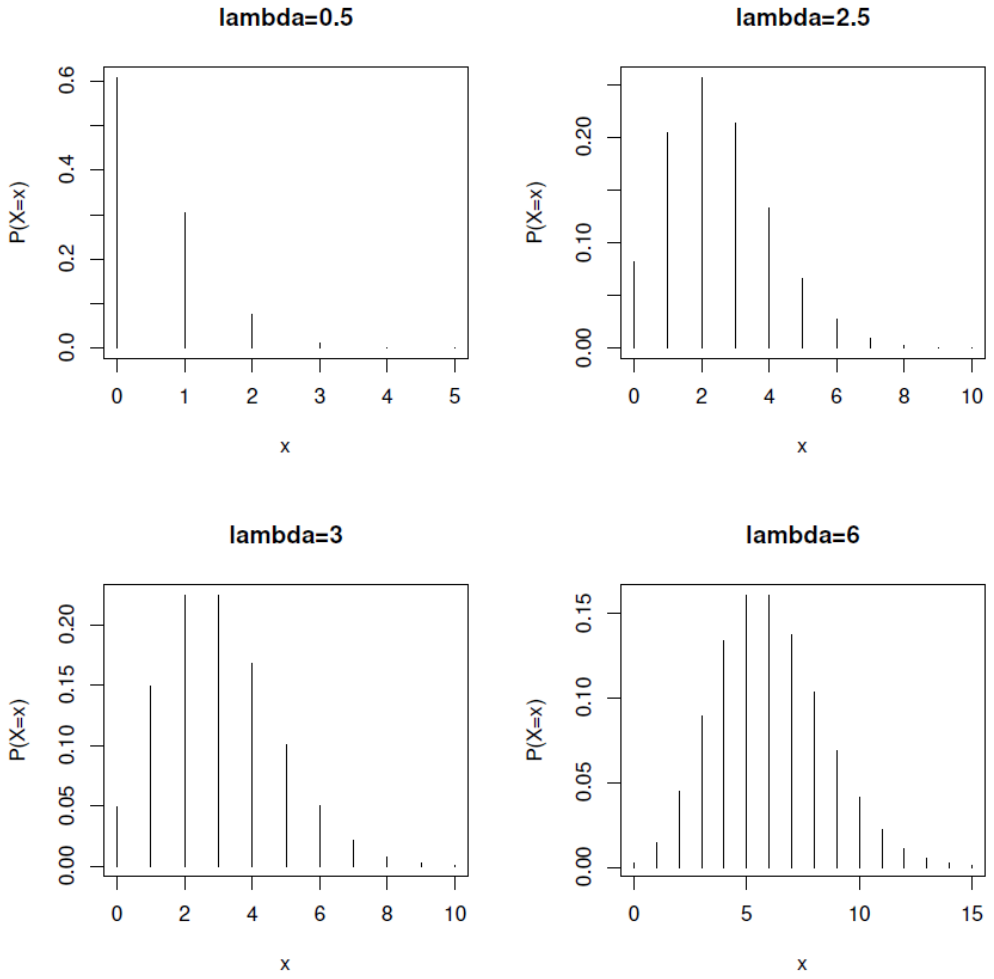
- Esempio: Consideriamo diversi valori se $\lambda = 0.5, 2.5, 3, 6$ le probabilità di Poisson per $x = 0, 1, \ldots, 10$ possono essere così valutate:

```
>par(mfrow=c(2,2))
>x<-0:5
>plot(x,dpois(x,lambda=0.5),
+xlab="x",ylab="P(X=x)",type="h",main="lambda=0.5")
>
>x<-0:10
>plot(x,dpois(x,lambda=2.5),
+xlab="x",ylab="P(X=x)",type="h",main="lambda=2.5")
>
>x<-0:10
>plot(x,dpois(x,lambda=3),
+xlab="x",ylab="P(X=x)",type="h",main="lambda=3")
>
>x<-0:15
>plot(x,dpois(x,lambda=6),
+xlab="x",ylab="P(X=x)",type="h",main="lambda=6")
```

Distribuzione di Poisson

- Esempio: Consideriamo diversi valori se $\lambda = 0.5,2.5,3,6$ le probabilità di Poisson per $x = 0, 1, \ldots, 10$ possono essere così valutate:

```
>par(mfrow=c(2,2))
>x<-0:5
>plot(x,dpois(x,lambda=0.5),
+xlab="x",ylab="P(X=x)",type="h",main="lambda=0.5")
>
>x<-0:10
>plot(x,dpois(x,lambda=2.5),
+xlab="x",ylab="P(X=x)",type="h",main="lambda=2.5")
>
>x<-0:10
>plot(x,dpois(x,lambda=3),
+xlab="x",ylab="P(X=x)",type="h",main="lambda=3")
>
>x<-0:15
>plot(x,dpois(x,lambda=6),
+xlab="x",ylab="P(X=x)",type="h",main="lambda=6")
```

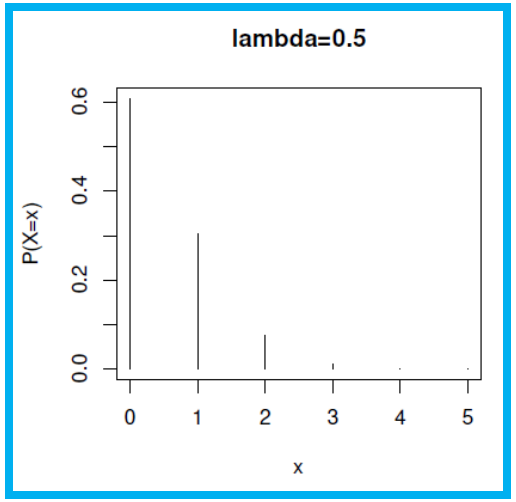


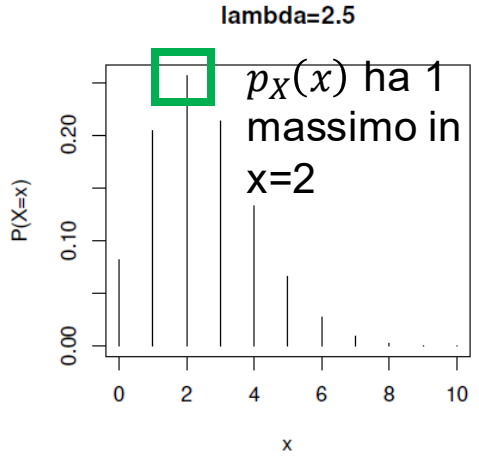
Distribuzione di Poisson

- Esempio: Consideriamo diversi valori se $\lambda = 0.5,2.5,3,6$ le probabilità di Poisson per $x = 0, 1, \ldots, 10$ possono essere così valutate:

```
>par(mfrow=c(2,2))
>x<-0:5
>plot(x,dpois(x,lambda=0.5),
+xlab="x",ylab="P(X=x)",type="h",main="lambda=0.5")
>
>x<-0:10
>plot(x,dpois(x,lambda=2.5),
+xlab="x",ylab="P(X=x)",type="h",main="lambda=2.5")
>
>x<-0:10
>plot(x,dpois(x,lambda=3),
+xlab="x",ylab="P(X=x)",type="h",main="lambda=3")
>
>x<-0:15
>plot(x,dpois(x,lambda=6),
+xlab="x",ylab="P(X=x)",type="h",main="lambda=6")
```

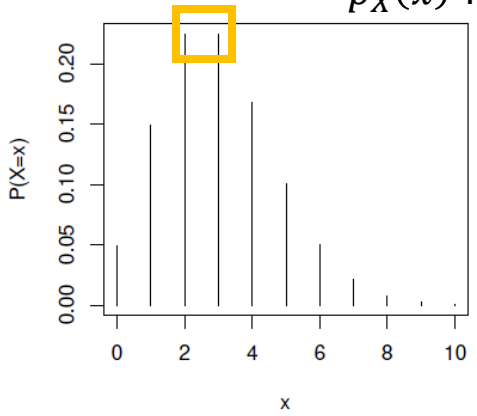
$p_X(x)$ è strettamente decrescente

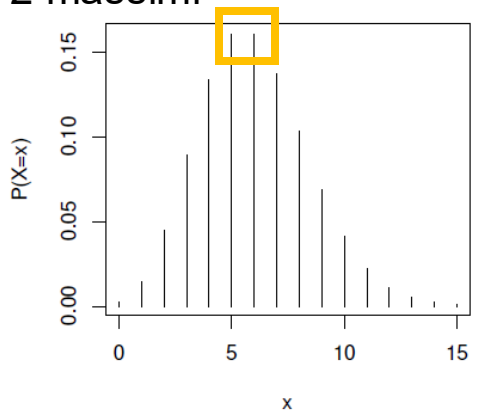




$p_X(x)$ ha 1 massimo in x=2

$p_X(x)$ ha 2 massimi





Distribuzione di Poisson

- Per il calcolo in R delle funzioni di distribuzione si usa

```
ppois(x, lambda, lower.tail = TRUE)
```

dove

  - **x** è il valore assunto (o i valori assunti) dalla variabile aleatoria di Poisson;
  - **lambda** è vettore dei valori medi (non negativi)
  - **lower.tail** se tale parametro è TRUE (caso di default) calcola $P(X \leq x)$, mentre se tale parametro è FALSE calcola $P(X > x)$

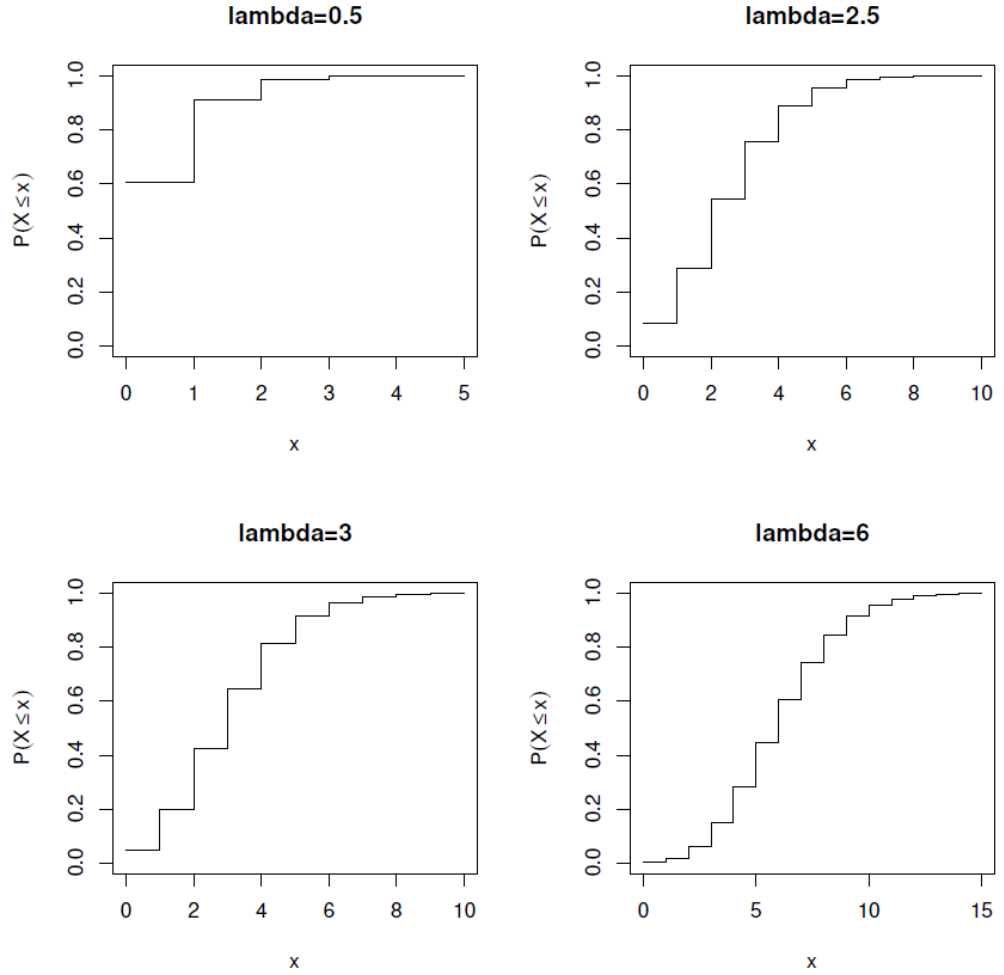
- Esempio: se $\lambda = 5$ le probabilità di Poisson per $x = 0, 1, \ldots, 8$ possono essere così valutate:

```
> x<-0:8
> ppois(x,lambda=0.5)
[1] 0.6065307 0.9097960 0.9856123 0.9982484 0.9998279 0.9999858
[7] 0.9999990 0.9999999 1.0000000
```

Distribuzione di Poisson

- Esempio: Consideriamo diversi valori se $\lambda = 0.5,2.5,3,6$ le funzioni di distribuzione di Poisson per $x = 0, 1, \ldots, 8$ possono essere così rappresentate:

```
>par(mfrow=c(2,2))
>x<-0:5
>plot(x,ppois(x,lambda=0.5),
+xlab="x",ylab=expression(P(X<=x)),ylim=c(0,1),type="s",
+main="lambda=0.5")
>
>x<-0:10
>plot(x,ppois(x,lambda=2.5),
+xlab="x",ylab=expression(P(X<=x)),ylim=c(0,1),type="s",
+main="lambda=2.5")
>
>x<-0:10
>plot(x,ppois(x,lambda=3),
+xlab="x",ylab=expression(P(X<=x)),ylim=c(0,1),type="s",
+main="lambda=3")
>
>x<-0:15
>plot(x,ppois(x,lambda=6),
+xlab="x",ylab=expression(P(X<=x)),ylim=c(0,1),type="s",
+main="lambda=6")
```



Quantili

- Per calcolare i quantili (percentili) della si utilizza la funzione basta aggiungere *n* ai quantili e alla simulazione della variabile binomiale negativa:

```
qpois(z, lambda)
```

dove

  - **z** è il valore assunto (o i valori assunti) dalle probabilità relative al percentile z · 100–esimo, cioè il vettore delle probabilità;
  - **lambda** è vettore dei valori medi (non negativi)

- Esempio: se $\lambda = 3$ i quartili Q_0, Q_1, Q_2, Q_3, Q_4 sono:

```
>z<-c(0,0.25,0.5,0.75,1)
> qpois(z,lambda=3)
[1]   0   2   3   4 Inf
```

$Q_0, = 0$ $Q_1 = 2$, $Q_2 = 3$, $Q_3 = 4$, $Q_4 = inf$

Quantili

- Esempio:
  - Generiamo una sequenza di 50 numeri pseudocasuali simulando una variabile aleatoria di Poisson con valore medio $\lambda = 3$:

```
> sim<-rpois(50,lambda=3)
> sim
[1] 1 1 4 0 3 0 2 1 5 3 2 1 5 2 2 2 5 1 4 2 1 4 2 3 5 4 3 4 0 3 3 1
[33] 7 2 1 2 2 2 3 4 2 1 2 5 1 2 1 4 5 3
> table(sim)
sim
 0  1  2  3  4  5  7
 3 11 14  8  7  6  1
> table(sim)/length(sim)
sim
   0    1    2    3    4    5    7
0.06 0.22 0.28 0.16 0.14 0.12 0.02
```

Frequenze assolute

Frequenze relative

Generazione di Osservazioni Sintetiche

- È possibile simulare in R la variabile aleatoria di Poisson generando una sequenza di numeri pseudocasuali mediante la funzione;

```
rpois(N,lambda)
```

dove

  - **N** è la lunghezza della sequenza da generare
  - **lambda** è vettore dei valori medi (non negativi)

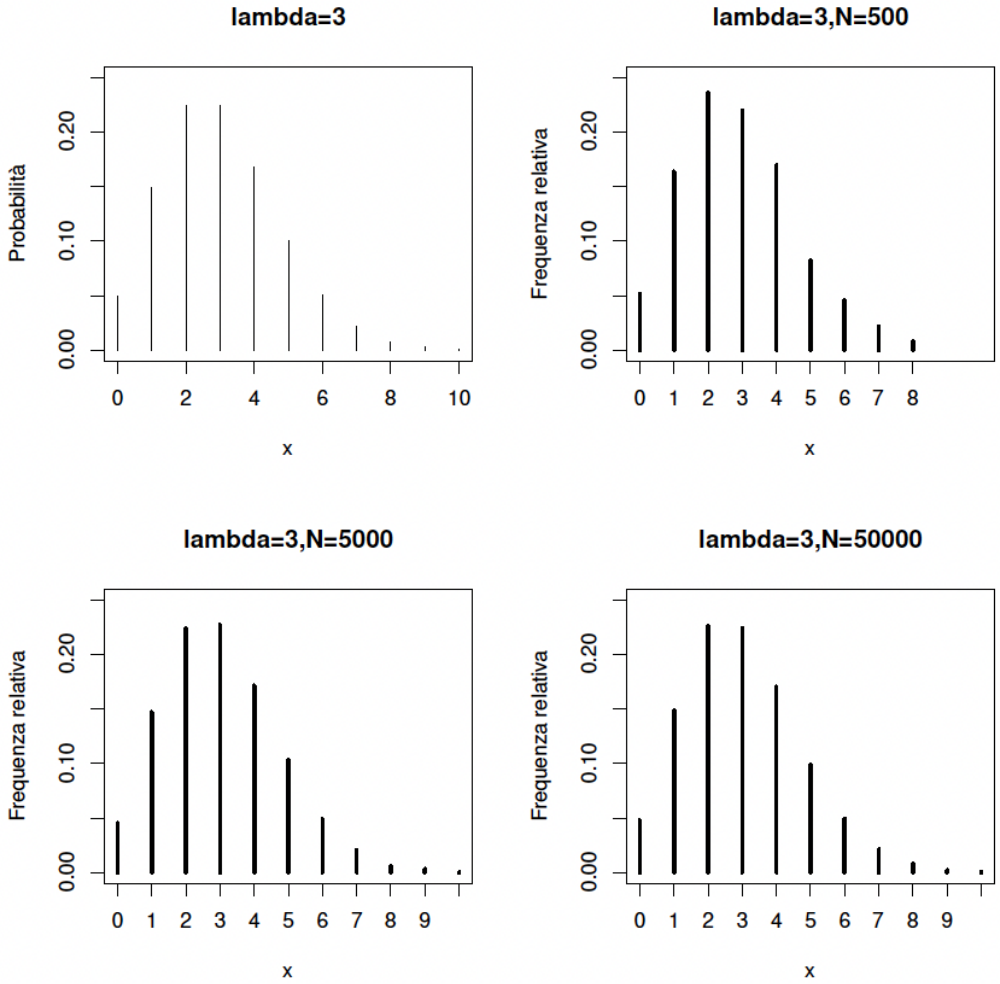
- Ad esempio, se desideriamo generare una sequenza di 50 numeri pseudocasuali simulando una variabile aleatoria di Poisson di valor medio λ = 3 si ha:

```
> sim<-rpois(50,lambda=3)
> sim
[1] 1 1 4 0 3 0 2 1 5 3 2 1 5 2 2 2 5 1 4 2 1 4 2 3 5 4 3 4 0 3 3 1
[33] 7 2 1 2 2 2 3 4 2 1 2 5 1 2 1 4 5 3
> table(sim)
sim
 0  1  2  3  4  5  7
 3 11 14  8  7  6  1
> table(sim)/length(sim)
sim
   0    1    2    3    4    5    7
0.06 0.22 0.28 0.16 0.14 0.12 0.02
```

Generazione di Osservazioni Sintetiche

- Confrontiamo la funzione di probabilità di Poisson teorica con quella simulata all'aumentare della lunghezza N = 500, 5000, 50000 della sequenza generata

```
>par(mfrow=c(2,2))
>x<-0:10
>plot(x,dpois(x,lambda=3),xlab="x",ylab="Probabilita' ",type="h",
+main="lambda=3",xlim=c(0,10),ylim=c(0,0.25))
>
>sim1<-rpois(500,lambda=3)
>plot(table(sim1)/length(sim1),xlab="x",type="h",
+ylab="Frequenza relativa",xlim=c(0,10),ylim=c(0,0.25),
+main="lambda=3,N=500")
>
>sim2<-rpois(5000,lambda=3)
>plot(table(sim2)/length(sim2),xlab="x",type="h",
+ylab="Frequenza relativa",xlim=c(0,10),ylim=c(0,0.25),
+main="lambda=3,N=5000")
>
>sim3<-rpois(50000,lambda=3)
>plot(table(sim3)/length(sim3),xlab="x",type="h",
+ylab="Frequenza relativa",xlim=c(0,10),ylim=c(0,0.25),
+main="lambda=3,N=50000")
```



Si nota che all'aumentare della lunghezza della sequenza generata il grafico delle frequenze relative si avvicina sempre di più al grafico della funzione di probabilità di Poisson.

STATISTICA E ANALISI DEI DATI

Distribuzione Ipergeometrica

Distribuzione Ipergeometrica

- La distribuzione Ipergeometrica interviene specificamente **nella descrizione di estrazioni senza reinserimento** oppure di estrazioni in blocco
- Si consideri l'esperimento che consiste nell'estrarre k biglie senza reinserimento da un'urna contenente $m + n$ biglie, di cui m sono bianche e n sono nere $(0 \leq k \leq m + n)$ e si consideri l'evento

$$E_r = \{r \text{ delle } k \text{ biglie estratte sono bianche}\} \quad (r = 0, 1, \ldots, k)$$

- Esempi:
  - Su 100 pezzi prodotti, 5 sono difettosi. Qual è la probabilità che in un campione di 10 pezzi ci siano almeno 2 difettosi?
  - In un mazzo di 52 carte con 13 cuori, qual è la probabilità di ottenere esattamente 3 cuori pescando 5 carte?
  - In una popolazione di 100 animali, 20 sono marcati. Qual è la probabilità che in un campione di 15 animali ci siano esattamente 4 marcati?
  - In una città di 10.000 abitanti, 3.000 preferiscono il prodotto A. Intervistando 100 persone, qual è la probabilità che almeno 40 preferiscano il prodotto A?
  - …

Distribuzione Ipergeometrica

- La distribuzione Ipergeometrica interviene specificamente **nella descrizione di estrazioni senza reinserimento** oppure di estrazioni in blocco
- Si consideri l'esperimento che consiste nell'estrarre k biglie senza reinserimento da un'urna contenente $m + n$ biglie, di cui m sono bianche e n sono nere $(0 \leq k \leq m + n)$ e si consideri l'evento

$$E_r = \{r \text{ delle } k \text{ biglie estratte sono bianche}\} \quad (r = 0, 1, \ldots, k)$$

Facendo ricorso alla definizione classica di probabilità si ha:

$$P(E_r) = \frac{\binom{m}{r}\binom{n}{k-r}}{\binom{m+n}{k}}$$

con $0 \leq r \leq m$ e $0 \leq k - r \leq n$, ossia:

$$\max(0, k - n) \leq r \leq \min(m, k)$$

Distribuzione Ipergeometrica

E_r = {r delle k biglie estratte sono bianche} $(r = 0, 1, \ldots, k)$

Facendo ricorso alla definizione classica di probabilità si ha:

È il numero di modi per scegliere le r biglie bianche dalle m disponibili

Numero di modi in cui si estrarre k delle $m + n$ biglie nell'urna

$$P(E_r) = \frac{\binom{m}{r}\binom{n}{k-r}}{\binom{m+n}{k}}$$

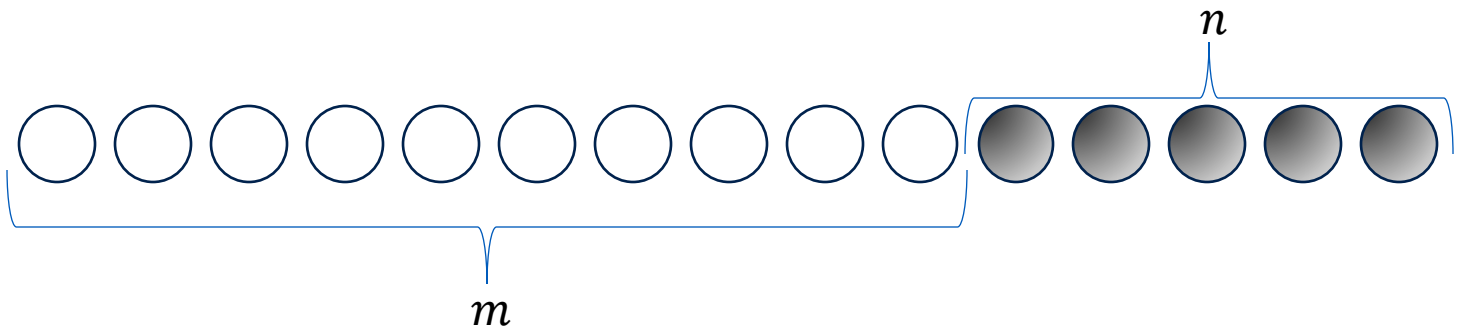
è il numero di modi per scegliere le restanti $k - r$ biglie (che devono essere nere) dalle n disponibili.

con $0 \leq r \leq m$ e $0 \leq k - r \leq n$, ossia:

$$max(0, k - n) \leq r \leq min(m, k)$$

Esempio: $m = 10, n = 5, k = 8$. Anche se volessi 0 biglie bianche, non potresti averle, perché $n = 5 < k = 8$.

Dovremmo per forza pescare almeno $k - n = 8 - 5 = 3$ biglie bianche. Quindi $r \geq 3$



Distribuzione Ipergeometrica

- Sia X la variabile aleatoria che descrive il numero di biglie bianche estratte senza reinserimento, risulta $X = r$ se e solo se si verifica l'evento E_r

Funzione di probabilità: $p_X(x) = \begin{cases} \frac{\binom{m}{x}\binom{n}{k-x}}{\binom{m+n}{k}} & \max(0, k-n) \leq r \leq \min(m, k) \\ 0 & altrimenti \end{cases}$

con n, m, k interi tali che $0 \leq k \leq m + n$ si dice avere distribuzione Ipergeometrica di parametri n, m, k

- Notazione:
  - $X \sim I(m, n, k)$ indicherà che X è una variabile aleatoria avente distribuzione Ipergeometrica di parametri m, n, k

Valore Atteso

- Possiamo definire **variabili aleatorie**:

$$X_i = \begin{cases} 1 & \textit{se la } i\textit{−esima biglia estratta è bianca} \\ 0 & \textit{altrimenti} \end{cases}$$

per $i = 1,2, \dots, k$.

  - Allora sia $X = X_1 + X_2 + \cdots + X_k$. Per la linearità del valore atteso: $E[X] = E\left[\sum_{i=1}^{k} X_i\right] = \sum_{i=1}^{k} E\left[X_i\right]$

- **Sia $N = m + n$, calcoliamo $E[X_i]$:**
  - Alla prima estrazione: $P(X_1 = 1) = \frac{m}{N}$
  - Alla seconda estrazione: $P(X_2 = 1) = \frac{m}{N}$ (per simmetria!)
  - Infatti, la probabilità che la seconda biglia sia bianca è:

$$\frac{m}{N} \cdot \frac{m-1}{N-1} + \frac{N-m}{N} \cdot \frac{m}{N-1} = \frac{m}{N}$$

  - Quindi **per ogni** i: $E[X_i] = \frac{m}{N}$
  - **Risultato finale:**

$$E[X] = \sum_{i=1}^{k} \frac{m}{N} = k \cdot \frac{m}{N} \quad \text{con} \quad N = n + m$$

Calcoliamo $P(X_2 = 1)$ usando la probabilità totale:

$$P(X_2 = 1) = P(X_1 = 1) \cdot P(X_2 = 1 \mid X_1 = 1) + P(X_1 = 0) \cdot P(X_2 = 1 \mid X_1 = 0)$$

Sostituiamo:

- $P(X_1 = 1) = \frac{m}{N}$
- $P(X_1 = 0) = \frac{N-m}{N}$
- $P(X_2 = 1 \mid X_1 = 1) = \frac{m-1}{N-1}$) se la prima era bianca, ne rimangono $m - 1$ su $N - 1$)
- $P(X_2 = 1 \mid X_1 = 0) = \frac{m}{N-1}$) se la prima era nera, rimangono tutte le m bianche su $N - 1$)

$$P(X_2 = 1) = \frac{m}{N} \cdot \frac{m-1}{N-1} + \frac{N-m}{N} \cdot \frac{m}{N-1}$$

Distribuzione Ipergeometrica

- Il valore medio e la varianza della distribuzione ipergeometrica risultano essere:

Valore atteso: $E(X) = k\frac{m}{m+n}$ **Varianza**: $Var(X) = k\frac{mn}{(m+n)^2}\frac{m+n-k}{m+n-1}$

- **Nota**: Se poniamo $p = \frac{m}{m+n}$ si nota che la media della distribuzione ipergeometrica coincide con la media di una variabile aleatoria binomiale $X \sim B(k,p)$ e la varianza della distribuzione ipergeometrica è $\frac{m+n-k}{m+n-1}$ volte la varianza della distribuzione binomiale

Distribuzione Ipergeometrica

- Il valore medio e la varianza della distribuzione ipergeometrica risultano essere:

Valore atteso: $E(X) = k\frac{m}{m+n}$ **Varianza**: $Var(X) = k\frac{mn}{(m+n)^2}\frac{m+n-k}{m+n-1}$

- **Nota**: Se poniamo $p = \frac{m}{m+n}$ si nota che la media della distribuzione ipergeometrica coincide con la media di una variabile aleatoria binomiale $X \sim B(k, p)$ e la varianza della distribuzione ipergeometrica è $\frac{m+n-k}{m+n-1}$ volte la varianza della distribuzione binomiale
- Per il calcolo in R della funzione di probabilità si utilizza la funzione:

```
dhyper(x, m, n, k)
```

dove

- $\boldsymbol{x}$ è il valore assunto (o i valori assunti) dalla variabile aleatoria ipergeometrica considerata;
- $\boldsymbol{m}$ indica il numero di palline bianche nell'urna;
- $\boldsymbol{n}$ indica il numero di palline nere nell'urna;
- $\boldsymbol{k}$ il numero di palline estratte dall'urna

Distribuzione Ipergeometrica

- Esempio: Consideriamo:
  - Il numero di palline bianche nell'urna è $m = 12$
  - Il numero di palline nere nell'urna è $n = 36$
  - Il numero di palline estratte dall'urna è $k = 5$,
  - $\max(0, k - n) = 0$ e $\min(m, k) = 5$
  - **x** rappresenta il numero di palline bianche estratte senza reinserimento da un'urna che contiene sia palline bianche che nere

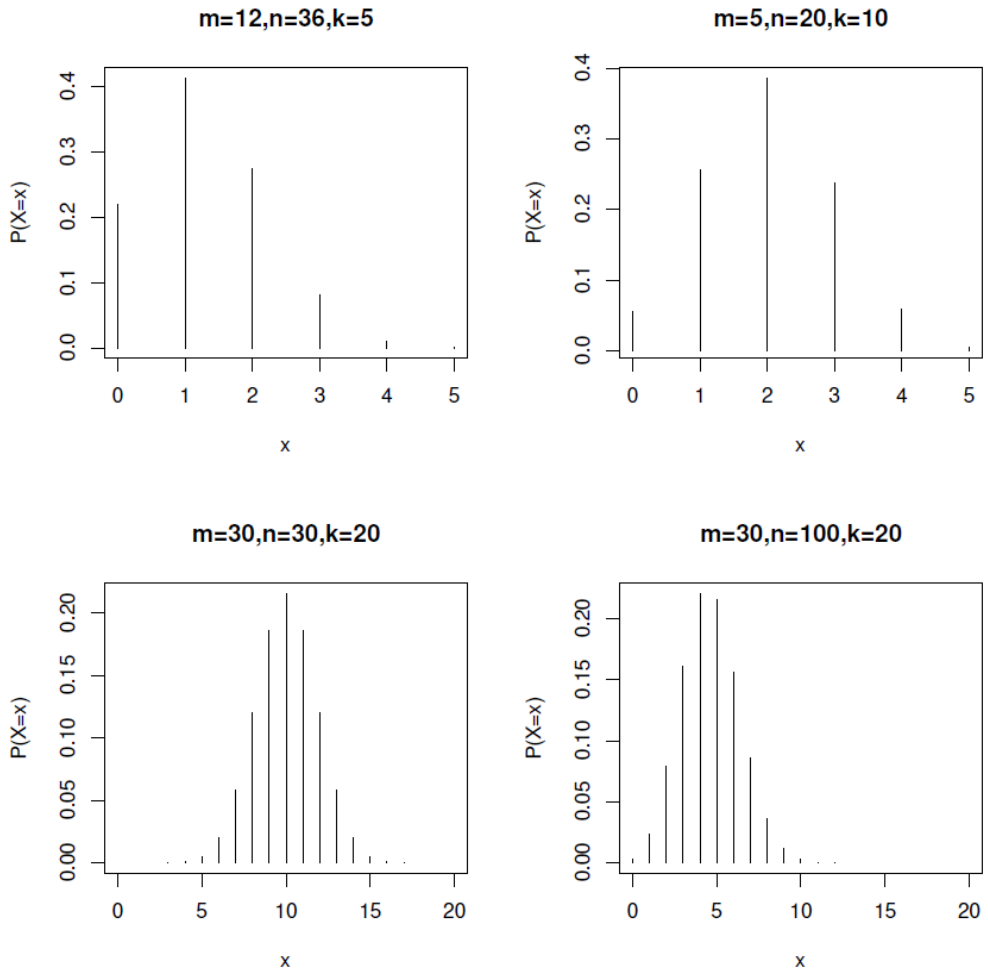
le probabilità ipergeometriche possono essere così valutate:

```
> x<-0:5
> dhyper(x,12,36,5)
[1] 0.2201665125 0.4128122109 0.2752081406 0.0809435708
[5] 0.0104070305 0.0004625347
```

Distribuzione Ipergeometrica

- Esempio: Visualizziamo le funzioni di probabilità ipergeometrica facendo variare k, m, n:

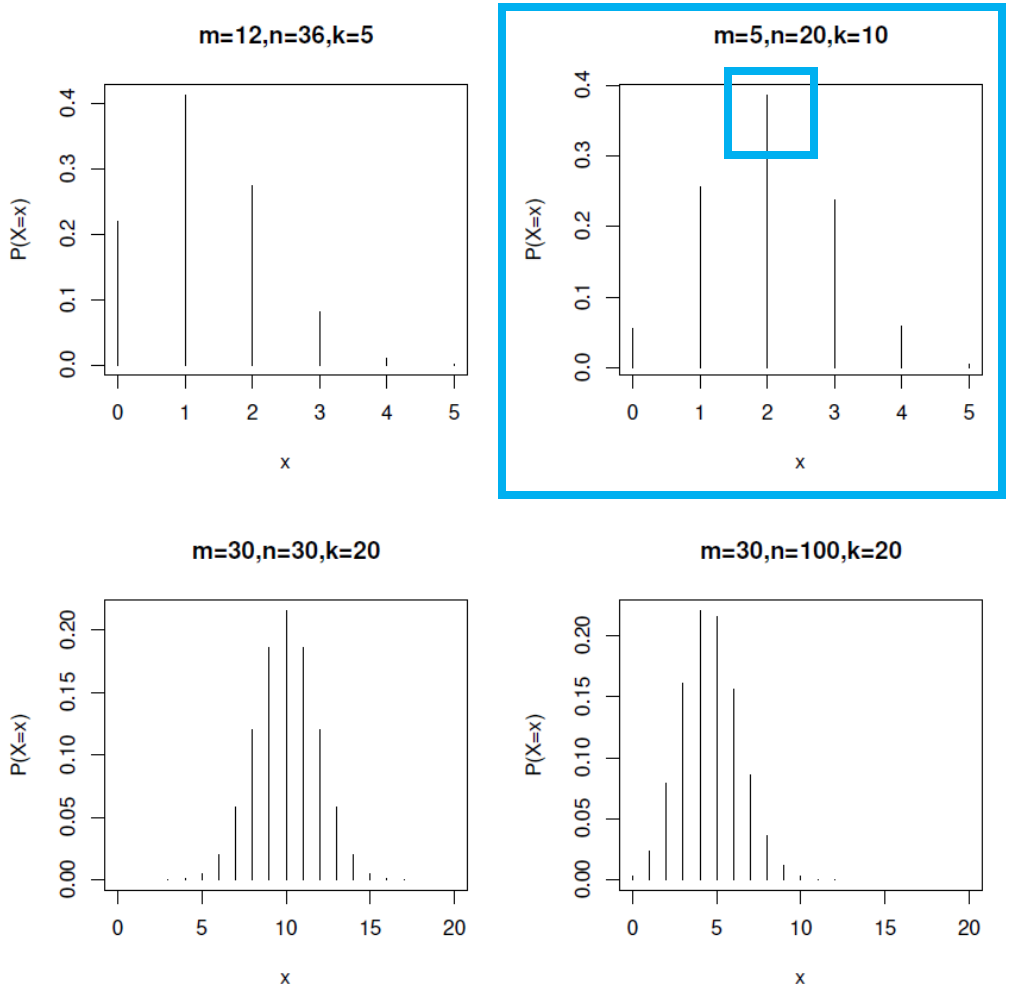
```
>par(mfrow=c(2,2))
>x<-0:5
>plot(x,dhyper(x,12,36,5),
+xlab="x",ylab="P(X=x)",type="h",
+main="m=12,n=36,k=5")
>
>x<-0:5
>plot(x,dhyper(x,5,20,10),
+xlab="x",ylab="P(X=x)",type="h",
+main="m=5,n=20,k=10")
>x<-0:20
>plot(x,dhyper(x,30,30,20),
+xlab="x",ylab="P(X=x)",type="h",
+main="m=30,n=30,k=20")
>
>x<-0:20
>plot(x,dhyper(x,30,100,20),
+xlab="x",ylab="P(X=x)",type="h",
+main="m=30,n=100,k=20")
```



Distribuzione Ipergeometrica

- Esempio: Visualizziamo le funzioni di probabilità ipergeometrica facendo variare k, m, n:

```
>par(mfrow=c(2,2))
>x<-0:5
>plot(x,dhyper(x,12,36,5),
+xlab="x",ylab="P(X=x)",type="h",
+main="m=12,n=36,k=5")
>
>x<-0:5
>plot(x,dhyper(x,5,20,10),
+xlab="x",ylab="P(X=x)",type="h",
+main="m=5,n=20,k=10")
>x<-0:20
>plot(x,dhyper(x,30,30,20),
+xlab="x",ylab="P(X=x)",type="h",
+main="m=30,n=30,k=20")
>
>x<-0:20
>plot(x,dhyper(x,30,100,20),
+xlab="x",ylab="P(X=x)",type="h",
+main="m=30,n=100,k=20")
```



Il caso più probabile è che 2 delle k biglie pescate siano bianche

Distribuzione Ipergeometrica

- Per il calcolo in R della funzione di distribuzione si utilizza la funzione:

```
phyper(x, m, n, k, lower.tail = TRUE)
```

dove

- **x** è il valore assunto (o i valori assunti) dalla variabile aleatoria ipergeometrica considerata;
- **m** indica il numero di palline bianche nell'urna;
- **n** indica il numero di palline nere nell'urna;
- **k** il numero di palline estratte dall'urna
- **lower.tail** se tale parametro è TRUE (caso di default) calcola $P(X \leq x)$, mentre se tale parametro è FALSE calcola $P(X > x)$

Distribuzione Ipergeometrica

- Esempio: Consideriamo l'esempio precedente con: il numero di palline bianche nell'urna è $m = 12$; il numero di palline nere nell'urna è $n = 36$; il numero di palline estratte dall'urna è $k = 5$ si ha che:

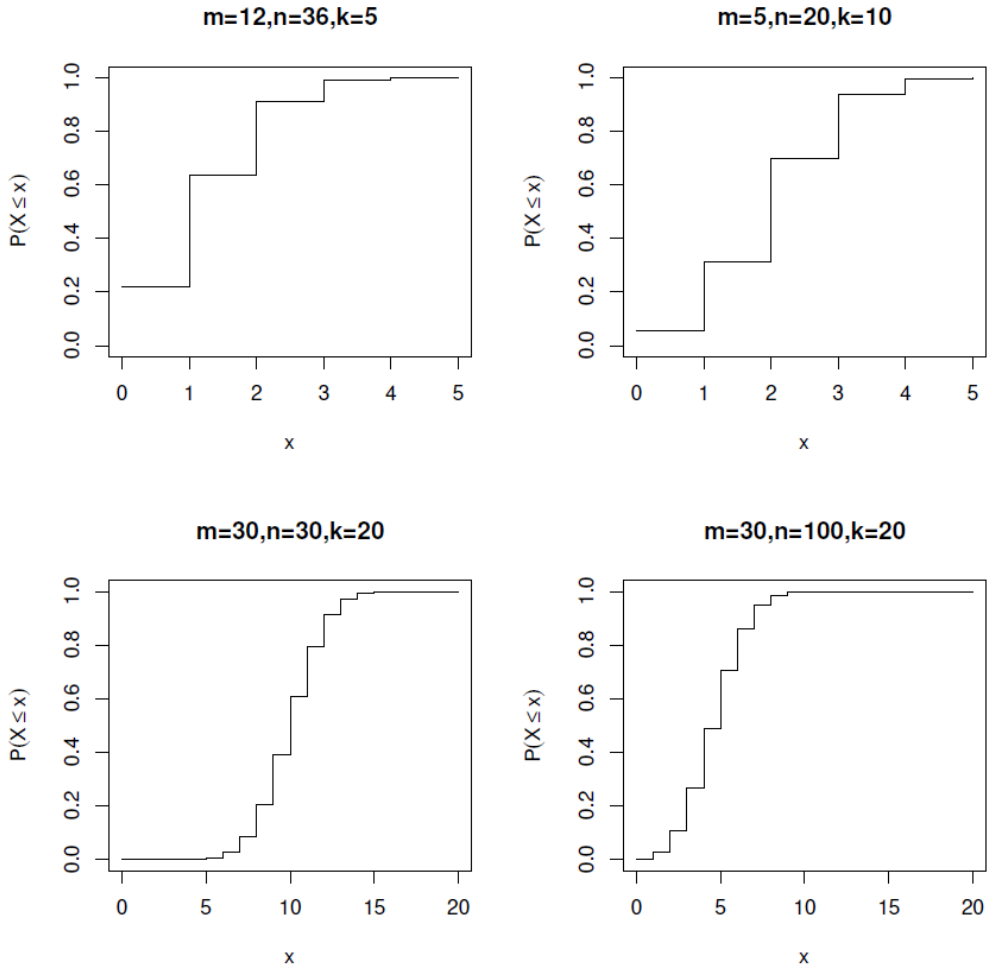
```
> x<-0:5
> phyper(x,12,36,5)
[1] 0.2201665 0.6329787 0.9081869 0.9891304 0.9995375 1.0000000
```

Distribuzione Ipergeometrica

- Esempio: Consideriamo l'esempio precedente con: il numero di palline bianche nell'urna è $m = 12$; il numero di palline nere nell'urna è $n = 36$; il numero di palline estratte dall'urna è $k = 5$ si ha che:

```
> x<-0:5
> phyper(x,12,36,5)
[1] 0.2201665 0.6329787 0.9081869 0.9891304 0.9995375 1.0000

>par(mfrow=c(2,2))
>x<-0:5
>plot(x,phyper(x,12,36,5),
+xlab="x",ylab=expression(P(X<=x)),ylim=c(0,1),type="s",
+main="m=12,n=36,k=5")
>
>x<-0:5
>plot(x,phyper(x,5,20,10),
+xlab="x",ylab=expression(P(X<=x)),ylim=c(0,1),type="s",
+main="m=5,n=20,k=10")
>
>x<-0:20
>plot(x,phyper(x,30,30,20),
+xlab="x",ylab=expression(P(X<=x)),ylim=c(0,1),type="s",
+main="m=30,n=30,k=20")
>
>x<-0:20
>plot(x,phyper(x,30,100,20),
+xlab="x",ylab=expression(P(X<=x)),ylim=c(0,1),type="s",
+main="m=30,n=100,k=20")
```



Quantili

- Per calcolare i quantili (percentili) della distribuzione Ipergeometrica si utilizza la funzione

```
qhyper(z, m, n, k)
```

dove

- **z** è il valore assunto (o i valori assunti) dalle probabilità relative al percentile z · 100–esimo;
- **m** indica il numero di palline bianche nell'urna;
- **n** indica il numero di palline nere nell'urna;
- **k** il numero di palline estratte dall'urna

- Esempio: se $m = 5, n = 20, k = 10$ i quartili Q_0, Q_1, Q_2, Q_3, Q_4 sono:

```
> z<-c(0,0.25,0.5,0.75,1)
> qhyper(z,5,20,10)
[1] 0 1 2 3 5
```

$Q_0, = 0$ $Q_1 = 1$, $Q_2 = 2$, $Q_3 = 3$, $Q_4 = 5$

Esempio

- Esempio:
  - Un produttore di computer decide di acquistare monitor da una nuova start-up che rivendica severi standard di controllo della qualità
  - Il produttore ordina 150 monitor e decide di accettare il lotto a condizione che un campione casuale di dimensione 25 non riveli monitor difettosi
  - Se il lotto di 150 monitor contiene 3 monitor difettosi, si desidera determinare la probabilità che il lotto venga accettato
- Denotiamo con X la variabile aleatoria che rappresenta il numero di monitor non difettosi nel campione. Si nota che $X \sim I(147, 3, 25)$. La probabilità richiesta è:

$$P(X = 25) = \frac{\binom{147}{25}\binom{3}{25-25}}{\binom{150}{25}} = 0.5764$$

```
> dhyper(25,147,3,25)
[1] 0.576365
```

Da Ipergeometrica a Binomiale

- Sia X la variabile aleatoria Ipergeometrica che rappresenta l'evento dell'estrazione di k biglie senza reinserimento da un'urna contenente $m + n$ biglie, di cui m sono bianche e n sono nere
- Se m e n divergono in modo che $\frac{m}{m+n}$ converga ad un valore $p \in (0,1)$ allora:

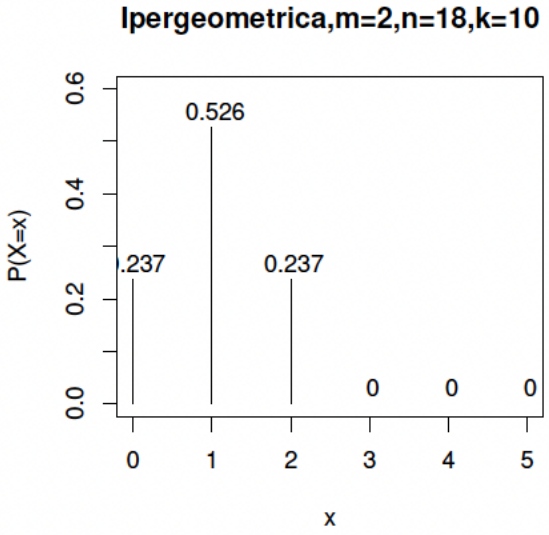
$$\lim_{\substack{m\to+\infty,\ m+n\to+\infty \\ m/(m+n)\to p}} p_X(x) = \binom{k}{x} p^x (1-p)^{k-x} \qquad (x = 0, 1, \ldots, k)$$

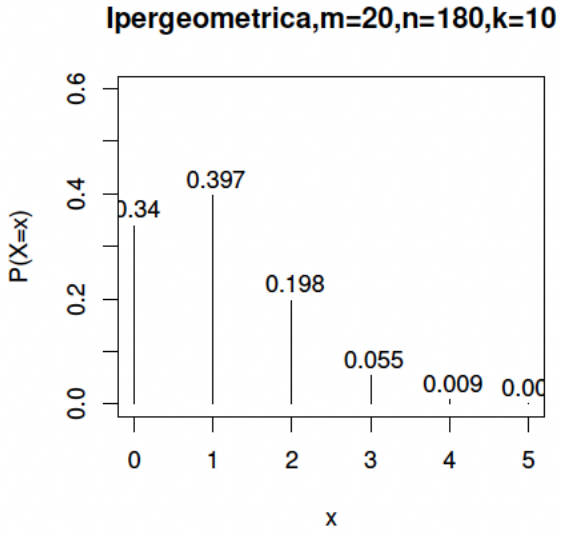
- Partendo da $P(E_r) = \frac{\binom{m}{r}\binom{n}{k-r}}{\binom{m+n}{k}}$, si ha che:
  - se il numero m delle biglie bianche e il numero $m + n$ di biglie presenti nell'urna sono entrambi sufficientemente elevati in modo tale che il loro rapporto sia una costante p,
  - Si ha che la probabilità che X delle k biglie estratte senza reinserimento siano bianche è approssimabile con la medesima probabilità relativa al caso di estrazioni con reinserimento

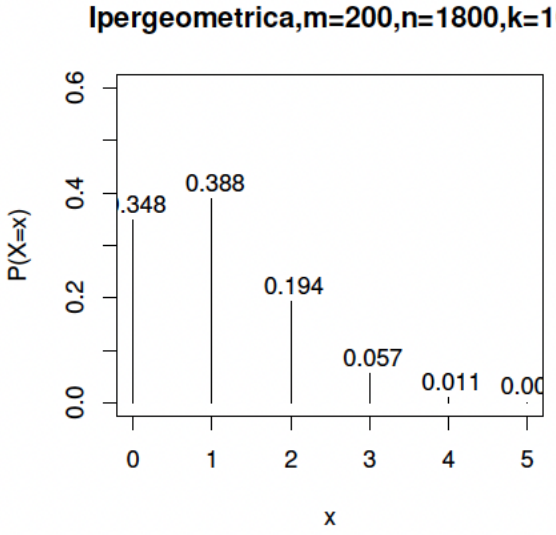
$$\frac{\binom{m}{x}\binom{n}{k-x}}{\binom{m+n}{k}} \simeq \binom{k}{x}\left(\frac{m}{m+n}\right)^x \left(1-\frac{m}{m+n}\right)^{k-x} \qquad (x = 0, 1, \ldots, k).$$

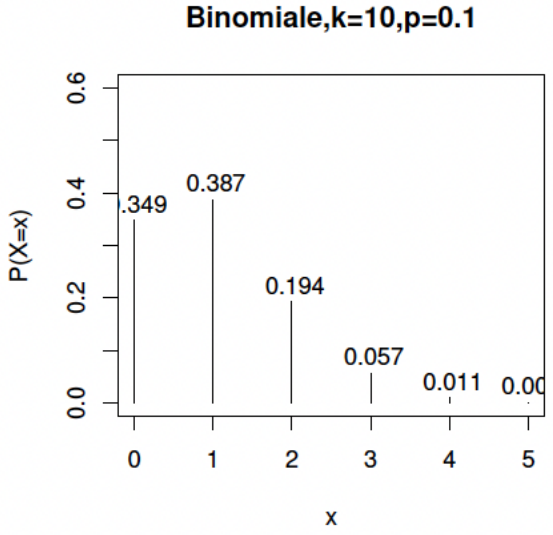
Da Ipergeometrica a Binomiale

```
>par(mfrow=c(2,2))
>x<-0:5
>plot(x,dhyper(x,2,18,10),
+xlab="x",ylab="P(X=x)",type="h",ylim=c(0,0.6),
+main="Ipergeometrica,m=2,n=18,k=10")
>y1<-round(dhyper(x,2,18,10),3)
>text(x+0.04,dhyper(x,2,18,10)+0.03,y1)
>
>x<-0:5
>plot(x,dhyper(x,20,180,10),
+xlab="x",ylab="P(X=x)",type="h",ylim=c(0,0.6),
+main="Ipergeometrica,m=20,n=180,k=10")
>y2<-round(dhyper(x,20,180,10),3)
>text(x+0.04,dhyper(x,20,180,10)+0.03,y2)
>
>x<-0:5
>plot(x,dhyper(x,200,1800,10),
+xlab="x",ylab="P(X=x)",type="h",ylim=c(0,0.6),
+main="Ipergeometrica,m=200,n=1800,k=10")
>y3<-round(dhyper(x,200,1800,10),3)
>text(x+0.04,dhyper(x,200,1800,10)+0.03,y3)
>
>x<-0:5
>plot(x,dbinom(x,size=10,prob=0.1),
+xlab="x",ylab="P(X=x)",type="h",ylim=c(0,0.6),
+main="Binomiale,k=10,p=0.1")
>y4<-round(dbinom(x,size=10,prob=0.1),3)
>text(x+0.04,dbinom(x,size=10,prob=0.1)+0.03,y4)
```









STATISTICA E ANALISI DEI DATI

Sommario

Funzioni di probabilità di variabili aleatorie discrete

Distribuzione	Notazione	Funzione di probabilità
Bernoulli	$X \sim \mathcal{B}(1,p)$	$p_X(x) = \begin{cases} 1-p, & x=0 \\ p, & x=1 \\ 0, & \text{altrimenti} \end{cases} \quad (0<p<1)$
Binomiale	$X \sim \mathcal{B}(n,p)$	$p_X(x) = \begin{cases} \binom{n}{x} p^x (1-p)^{n-x}, & x=0,1,\ldots,n \\ 0, & \text{altrimenti} \end{cases} \quad (n=1,2,\ldots;\ 0<p<1)$
Geometrica	$X \sim \mathcal{BN}(1,p)$	$p_X(x) = \begin{cases} p\,(1-p)^x, & x=0,1,\ldots \\ 0, & \text{altrimenti} \end{cases} \quad (0<p<1)$
Geometrica modificata	$X \sim \mathcal{BN}^*(1,p)$	$p_X(x) = \begin{cases} p\,(1-p)^{x-1}, & x=1,2,\ldots \\ 0, & \text{altrimenti} \end{cases} \quad (0<p<1)$
Binomiale negativa	$X \sim \mathcal{BN}(n,p)$	$p_X(x) = \begin{cases} \binom{n+x-1}{x} p^n (1-p)^x, & x=0,1,\ldots \\ 0, & \text{altrimenti} \end{cases} \quad (n=1,2,\ldots;\ 0<p<1)$
Binomiale negativa modificata	$X \sim \mathcal{BN}^*(n,p)$	$p_X(x) = \begin{cases} \binom{x-1}{n-1} p^n (1-p)^{x-n}, & x=n,n+1,\ldots \\ 0, & \text{altrimenti} \end{cases} \quad (n=1,2,\ldots;\ 0<p<1)$
Poisson	$X \sim \mathcal{P}(\lambda)$	$p_X(x) = \begin{cases} \frac{\lambda^x}{x!} e^{-\lambda}, & x=0,1,\ldots \\ 0, & \text{altrimenti} \end{cases} \quad (\lambda>0)$
Ipergeometrica	$X \sim \mathcal{I}(n,m,k)$	$p_X(x) = \begin{cases} \dfrac{\binom{m}{x}\binom{n}{k-x}}{\binom{m+n}{k}}, & x \geq \max\{0,k-n\} \\ & x \leq \min\{m,k\} \\ 0, & \text{altrimenti} \end{cases} \quad (0 \leq k \leq m+n)$

Valori medi, varianze e coefficienti di variazione

Nome	$E(X)$	$\mathrm{Var}(X)$	$\mathrm{CV}(X)$
Bernoulli	p	$p(1-p)$	$\sqrt{\frac{1-p}{p}}$
Binomiale	np	$np(1-p)$	$\sqrt{\frac{1-p}{np}}$
Geometrica	$(1-p)/p$	$(1-p)/p^2$	$1/\sqrt{1-p}$
Geometrica modificata	$1/p$	$(1-p)/p^2$	$\sqrt{1-p}$
Binomiale negativa	$n(1-p)/p$	$n(1-p)/p^2$	$1/\sqrt{(1-p)/n}$
Binomiale negativa modificata	n/p	$n(1-p)/p^2$	$\sqrt{(1-p)/n}$
Poisson	λ	λ	$1/\sqrt{\lambda}$
Ipergeometrica	$k\frac{m}{m+n}$	$k\frac{mn}{(m+n)^2}\frac{m+n-k}{m+n-1}$	$\sqrt{\frac{n(m+n-k)}{k\,m\,(m+n-1)}}$

Funzioni in R

Nome	Probabilità Distribuzione	Quantili	Simulazione
Bernoulli	dbinom(x,1,prob) pbinom(x,1,prob)	qbinom(z,1,prob)	rbinom(N,1,prob)
Binomiale	dbinom(x,size,prob) pbinom(x,size,prob)	qbinom(z,size,prob)	rbinom(N,size,prob)
Geometrica	dgeom(x, prob) pgeom(x, prob)	qgeom(z, prob)	rgeom(N, prob)
Geometrica modificata	dgeom(x-1, prob) pgeom(x-1, prob)	qgeom(z, prob)+1	rgeom(N, prob)+1
Binomiale negativa	dnbinom(x, size, prob) pnbinom(x, size, prob)	qnbinom(z, size, prob)	rnbinom(x, size, prob)
Binomiale negativa modificata	dnbinom(x-n, size, prob) pnbinom(x-n, size, prob)	qnbinom(z, size, prob)+n	rnbinom(x, size, prob)+n
Poisson	dpois(x,lambda) ppois(x,lambda)	qpois(z,lambda)	rpois(N,lambda)
Ipergeometrica	dhyper(x, m, n, k) phyper(x, m, n, k)	qhyper(z, m, n, k)	rhyper(N, m, n, k)