5.2 Distribuzioni discrete

In questa sezione si introducono le variabili casuali, discrete e continue. A partire dalle variabili casuali discrete, si mostrano alcune distribuzioni discrete di probabilità.

5.2.1 Variabili aleatorie

Definizione. Una *variabile aleatoria (o casuale)* X è una funzione

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

che associa ad ogni elemento $w \in S$ un numero $X(w) = x$.

Si indicano solitamente le variabili aleatorie con lettere maiuscole e il loro valore (osservato) con lettere minuscole.

Si dice *supporto di X* l'insieme di tutti i possibili valori casuali di un dato esperimento e si indica con S_X.

Esempio 5.18 *Consideriamo l'esperimento di lanciare una moneta due volte.*

Soluzione. $S = \{TT, TC, CT, CC\}$ e sia X la variabile casuale: $X(w) =$\{numero di C\}. Si ha: $X(TT) = 0, X(TC) = 1, X(CT) = 1, X(CC) = 2$. Il supporto è $S_X = \{0, 1, 2\}$ $\subset R$

Esempio 5.19 *Consideriamo ora il lancio ripetuto di un dado finché non si ottiene un* 6.

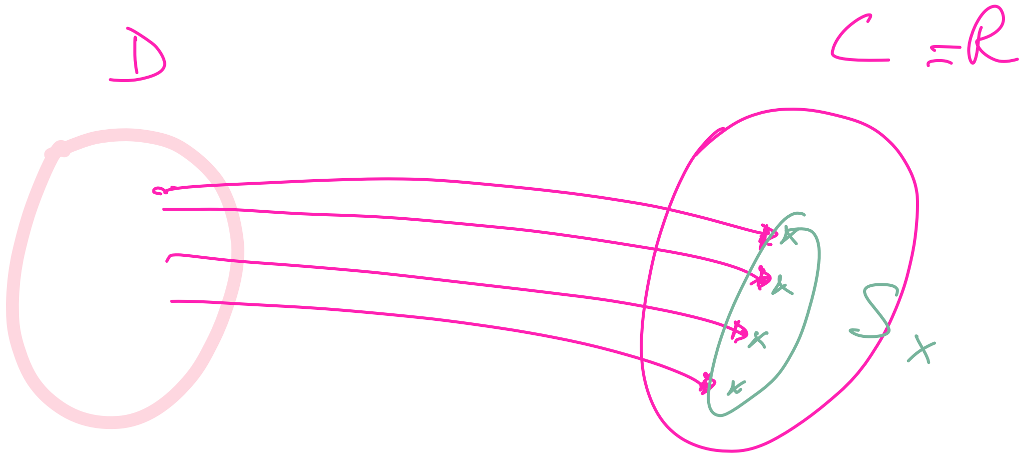
spazio campione

Soluzione. In questo caso, $S_X = \{6, \{16\}, \{26\}, \{36\}, \{46\}, \{56\}, \ldots\}$ e se si definisce $Y =$\{numero di lanci prima di ottenere 6\} si ha che $S_Y = \{0, 1 \ldots\}$.

Esempio 5.20 *Consideriamo come esperimento l'arrivo di un cliente in banca.*

Soluzione. Sia $Z =$\{tempo di attesa dell'arrivo del cliente\}. In questo caso $S_Z = (0, \infty)$. $[0, T] \subset R$

Nel caso delle variabili X e Y il supporto è costituito da un insieme numerabile (contabile) di elementi; nel primo caso l'insieme è costituito da un numero finito, nel secondo caso da un numero infinito di elementi. Queste sono *variabili aleatorie discrete*. Il supporto della variabile Z invece è un intervallo continuo, un insieme quindi non contabile. La variabile Z è una *variabile aleatoria continua*.



$S_x \subset R$

5.2.2 Funzione di densità di probabilità

Poiché il supporto S_X di una variabile aleatoria discreta X è un insieme numerabile, è possibile definire una funzione f_X, detta *funzione di massa o di probabilità di X (Probability Mass Function PMF)*

$x \in S_X$

$$\boxed{f_X : S_X \to [0,1], \quad f_X(x) = P(X = x), \quad x \in S_X.}$$

La PMF possiede alcune proprietà che derivano dal fatto che il suo valore è una probabilità.

1. $f_X(x) \geq 0 \ \forall x \in S_X$,
2. $\sum_{x \in S_X} f_X(x) = 1$,
3. $P(X \in A) = \sum_{x \in A} f_X(x)$ per ogni evento $A \subset S_X$.

Esempio 5.21 *Lancio di una moneta due volte.*

Soluzione. Lo spazio dei campioni è $S = \{TT, TC, CT, CC\}$. Considero X ={numero di C}. Allora $S_X = \{0, 1, 2\}$. $f_X(0) = 1/4, f_X(1) = 1/2, f_X(2) = 1/4$.

Associati alla PMF si hanno i seguenti valori:

- **media**

$E[X]$

$$\boxed{\mu = \mathbf{E}X = \sum_{x \in S_X} x f_X(x)}$$

(supponendo che la serie infinita $\sum |x| f_X(x)$ sia convergente),

- **varianza**

$$\boxed{\sigma^2 = \sum_{x \in S_X} (x - \mu)^2 f_X(x),}$$

- **deviazione standard** $\boxed{\sigma = \sqrt{\sigma^2}}$.

Si definisce inoltre la funzione generatrice dei momenti:

Definizione. Data una variabile aleatoria X discreta con PMF $f_X(x)$, la sua *funzione generatrice dei momenti* (MGF) è definita da:

$$M_X(t) = \sum_{x \in S} e^{tx} f_X(x),$$

supposto che la serie (quando è infinita) sia convergente per $-\epsilon < t < \epsilon$, per qualche $\epsilon > 0$.

Si ha in particolare che $M_X(0) = 1$.

Lancio moneta 2 volte. (esempio 5.18)

$X = \{\text{numero di croci}\}$

f_X ? $\qquad f_X : \boxed{S_X} \to [0,1]$

$$f_X : \begin{array}{l} 0 \longrightarrow \\ 1 \longrightarrow \\ 2 \longrightarrow \end{array}$$

$\underline{S_X = \{0, 1, 2\}}$

$S = \{TT, TC, CT, CC\}$

$X(TT) = 0. \qquad X(TC) = 1.$

$X(CT) = 1. \qquad X(CC) = 2.$

PMF
$$\begin{cases} f_X(0) = P(X=0) = \frac{1}{4} = 0.25 \\ f_X(1) = P(X=1) = \frac{1}{2} = 0.5 \\ f_X(2) = P(X=2) = \frac{1}{4} = 0.25 \end{cases}$$

$$\mu = \sum_{x \in S_X} x\, f_X(x) = \sum_{x=0}^{2} x\, f_X(x)$$

$S_X = \{0, 1, 2\}$

$x = 0 \longrightarrow 0\, f_X(0) = 0 \cdot \frac{1}{4} = 0$

$x = 1 \longrightarrow 1 \cdot f_X(1) = 1 \cdot \frac{1}{2} = \frac{1}{2}$

$x = 2 \longrightarrow 2 \cdot f_X(2) = 2 \cdot \frac{1}{4} = \frac{1}{2}$

$\rightarrow \quad \mu = 0 + \frac{1}{2} + \frac{1}{2} = 1$

$$\sigma^2 = \sum_{x \in S_X} \underline{(x-\mu)^2 f_X(x)} = \frac{1}{4} + 0 + \frac{1}{4} = \frac{1}{2}$$

$x = 0 \longrightarrow (0 - 1)^2 f_X(0) = 1 \cdot \frac{1}{4} = \frac{1}{4}$

$x = 1 \longrightarrow (1 - 1)^2 f_X(1) = 0 \cdot \frac{1}{2} = 0$

$x = 2 \longrightarrow (2 - 1)^2 f_X(2) = 1 \cdot \frac{1}{4} = \frac{1}{4}$

Esempio 5.22 *Calcolo di μ rispetto all'esperimento precedente.*

Soluzione.

$$\mu = 0 \cdot 1/4 + 1 \cdot 1/2 + 2 \cdot 1/4 = 1,$$
$$\sigma^2 = (-1)^2 \cdot 1/4 + 0 \cdot 1/2 + 1^2 \cdot 1/4 = 0.5.$$

Definizione. Si dice *funzione di ripartizione o di distribuzione di una variabile aleatoria discreta (Cumulative Distribution Function CDF)* $F_X(t)$ la funzione:

$$\boxed{F_X(t) = P(X \le t), \quad -\infty < t < \infty}$$

$F_X : R \to [0,1]$

La CDF soddisfa le seguenti proprietà:

1. $F_X(t)$ è non decrescente
2. $F_X(t)$ è continua a destra, cioè: $lim_{t \to a^+} F_X(t) = F_X(a)$
3. vale: $lim_{t \to -\infty} F_X(t) = 0$ e $lim_{t \to +\infty} F_X(t) = 1$.

Nel seguito, consideriamo particolari modelli o distribuzioni utilizzate in pratica.

5.2.3 Distribuzione uniforme

La distribuzione discreta più comune è la cosiddetta *distribuzione discreta uniforme* in cui la PMF è definita:

$$\boxed{f_X(x) = \frac{1}{n}, x = 1, \ldots n.}$$

Si scrive $X \simeq$ `disunif(n)`.

Per la media μ vale:

$$\mu = \sum_{x=1}^{n} x f_X(x) = \sum_{x=1}^{n} x \cdot 1/n = \frac{n(n+1)}{2n} = \frac{n+1}{2}$$

Per la varianza σ^2 calcolo innanzitutto:

$$\sum_{x=1}^{n} x^2 f_X(x) = 1/n \sum_{x=1}^{n} x^2 = \frac{n(n+1)(2n+1)}{6n} = \frac{(n+1)(2n+1)}{6}$$

quindi:

$$\sigma^2 = \sum_{x=1}^{n} x^2 f_X(x) - \mu^2 = \frac{(n+1)(2n+1)}{6} - \frac{n+1}{2} = \frac{n^2-1}{12}.$$

Esempio 5.23 *Lancio un dado e sia X ={il risultato del lancio}.*

Soluzione. Allora $S_X = 1, \ldots 6$ e $f_X(x) = 1/6\,\forall x \in [1,6]$. Quindi

$$\mu = 1/6 + 2/6 + 3/6 + 4/6 + 5/6 + 6/6 = 21/6 = 3.5,$$

$$\sigma^2 = (6^2 - 1)/12 = 35/12.$$

In R, per simulare una variabile aleatoria discreta da una distribuzione uniforme si può usare la funzione `sample`, indicando con `replace=TRUE` che ci possono essere valori ripetuti.

- estrazione di 5 numeri casuali interi in [0, 100].

```
> sample(1:100, size = 5, replace = TRUE)

## [1] 61 81  3 60 65
```

- Lancio di un dado 10 volte.

```
> sample(1:6, size = 10, replace = TRUE)

##  [1] 1 3 4 1 2 2 3 5 3 3
```

- lancio di una moneta 10 volte

```
> sample(c('T', 'C'), size = 10, replace = TRUE)

##  [1] "T" "T" "T" "C" "T" "T" "T" "T" "T" "T"
```

5.2.4 Distribuzione binomiale

La *distribuzione binomiale* si basa sul processo di Bernoulli, che è un esperimento casuale che ha solo due possibili risultati: successo (S) o insuccesso (I). Sia X la variabile aleatoria tale che: $\{X = 0$ se il risultato è S, $X = 1$ se il risultato è $I\}$. Se indichiamo con p la probabilità di successo del singolo processo, la PMF è:

$$f_X(x) = p^x(1-p)^{1-x}, \;\; x = 0, 1.$$

Si ha che $\mu = p$ e $\sigma^2 = np(1-p)$.

Il modello binomiale si basa su 3 principi:

- È costituito da n processi di Bernoulli,
- i processi sono fra loro indipendenti,
- ogni processo ha la stessa probabilità p di successo.

Se X è la variabile aleatoria che conta i successi negli n processi, allora la PMF di X è:

$$f_X(x) = \binom{n}{x} p^x (1-p)^{n-x}, \quad x = 0, 1, \ldots n.$$

$$\binom{n}{x} = \frac{n!}{x!(n-x)!}$$

Allora X ha una *distribuzione binomiale*, $X \simeq$ `binom(size=n,prob=p)`. Si ha che la media della distribuzione è $\mu = np$ e la varianza $\sigma^2 = n(n-1)p^2$.

Esempio 5.24 *In una famiglia di 5 figli, sia X =\{numero di femmine\}. Se supponiamo che la nascita di un maschio abbia probabilità 49% e quella di una femmina 51%, qual è la probabilità $P(X = 2)$?*

Soluzione. ~~In questo caso, $X \simeq$ `binom(size=5,prob=0.51)`~~. Quindi:

$$f_X(2) = \binom{5}{2} 0.51^2 0.49^3 = 0.306...$$

In R ci sono, nel pacchetto `stats`, R Core Team [2015], delle funzioni, relative a diverse distribuzioni, che hanno tutte una sintassi simile, in cui la lettera iniziale contraddistingue il tipo di output della funzione per quella particolare distribuzione. Per esempio, nel caso della distribuzione binomiale si ha:

- **d**binom che ritorna la funzione di densità,
- **p**binom che ritorna la funzione di ripartizione o distribuzione,
- **q**binom che ritorna i quantili,
- **r**binom che genera valori casuali appartenenti alla distribuzione.

```
> library(prob)
> dbinom(0:5, size = 5, prob = 0.51)

## [1] 0.02824752 0.14700243 0.30600505 0.31849505
0.16574742 0.03450253
```

da cui si ricava che $P(X = 2) = 0.306\ldots$. Oppure nel pacchetto `distr`, Ruckdeschel et al. [2006] e Ruckdeschel and Kohl [2014], c'è un approccio ad oggetti delle distribuzioni.

Esempio 5.4

$P(X=2)$

$$f_X(x) = \binom{n}{x} p^x (1-p)^{n-x}$$ ← PMF

→ $p = 0.51$ (probabilità di successo → nascita di una femmina)

→ $n = 5$ (numero processi di Bernoulli)

$$f_X(2) = \binom{5}{2} 0.51^2 (1-0.51)^{5-2}$$

Stesso testo, $Y = \{$numero di maschi$\}$

$P(Y=3)$ nascono 3 maschi

$p = 0.49$ $\qquad y = 3$

$n = 5$

$$f_Y(3) = \binom{5}{3} 0.49^3 (1-0.49)^{5-3}$$

$$f_Y(3) = \binom{n}{3} p^3 (1-p)^{n-3}$$

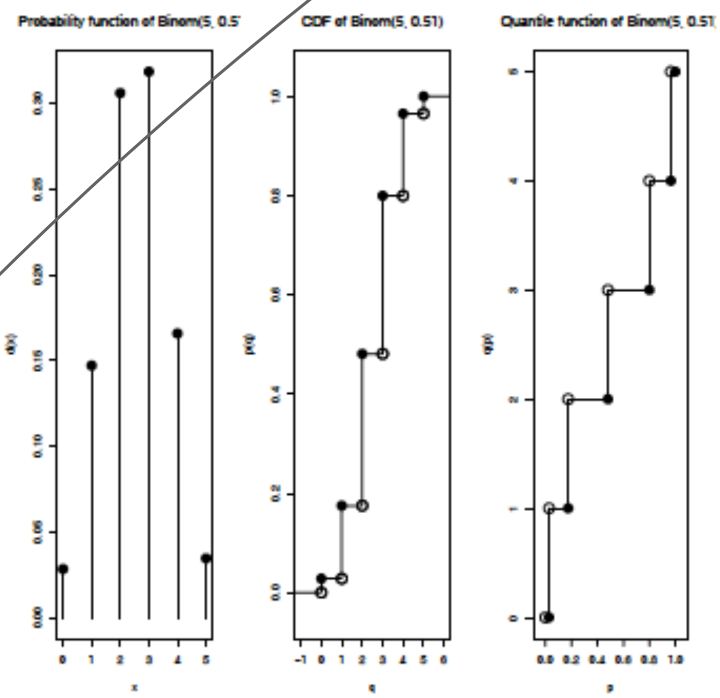
```
> library(distr)
> Binom(size = 5, prob = 0.51)

## Distribution Object of Class: Binom
##  size: 5
##  prob: 0.51
```

È possibile anche usare la funzione `plot` dello stesso pacchetto:

```
> library(distr)
> plot(Binom(size = 5, prob = 0.51))

## NULL
```



L'analogo delle funzioni `dbinom` e `pbinom` sono rispettivamente `d(X)` e `p(X)` se X è il nome che si è dato alla distribuzione:

```
> library(distr)
> X <- Binom(size = 5, prob = 0.51)
> d(X)(1)

## [1] 0.1470024

> p(X)(2)

## [1] 0.481255
```

Calcolare la probabilità di ottenere esattamente 3 volte 6.

Esempio 5.25 *Lancio di 5 dadi contemporaneamente. Sia X={numero di 6 che escono}. ~~Calcolare la probabilità di ottenere un numero di 6 nell'intervallo [3, 5].~~*

Soluzione. Ogni dado lanciato rispetto ad avere come risultato 6 è un processo di Bernoulli con probabilità $p = 1/6$, quindi $X \simeq$ `binom(size=5,prob=1/6)`. Dobbiamo calcolare $P(3 \leq X \leq 5)$. Quindi:

$$P(3 \leq X \leq 5) = \sum_{x=7}^{9} \binom{5}{2} (1/6)^x (5/6)^{5-x}.$$

Oppure posso scrivere:

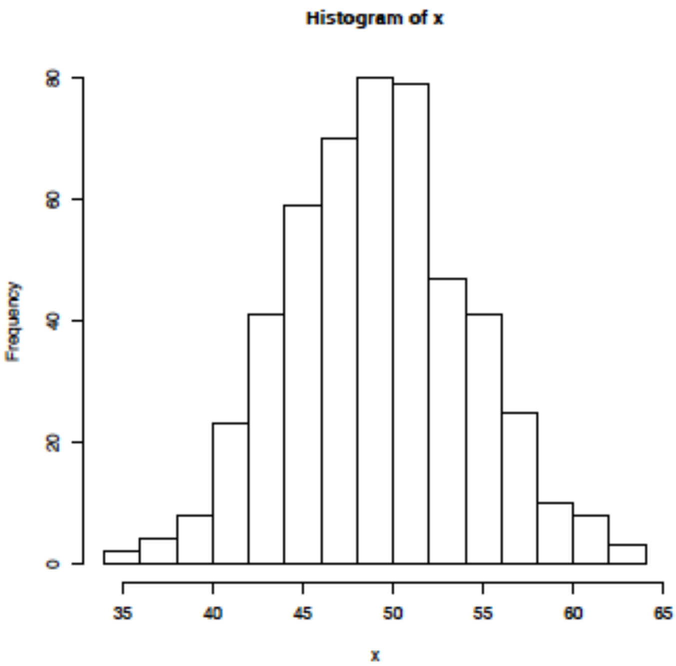
$$P(3 \leq X \leq 5) = P(X \leq 5) - P(X \leq 2) = F_X(5) - F_X(3).$$

```
> pbinom(5, size = 5, prob = 1/6) - pbinom(2, size = 5,
+                                            prob = 1/6)

## [1] 0.03549383
```

Per disegnare l'istogramma della distribuzione:

```
>  x <- rbinom(500, size = 100 , prob = 1/2)
> hist(x, breaks = 20)
```



X variabile aleatoria con distribuzione binomiale

$p = \frac{1}{6}$ $\quad n = 5 \rightarrow$ numero processi di Bernoulli

$\downarrow$

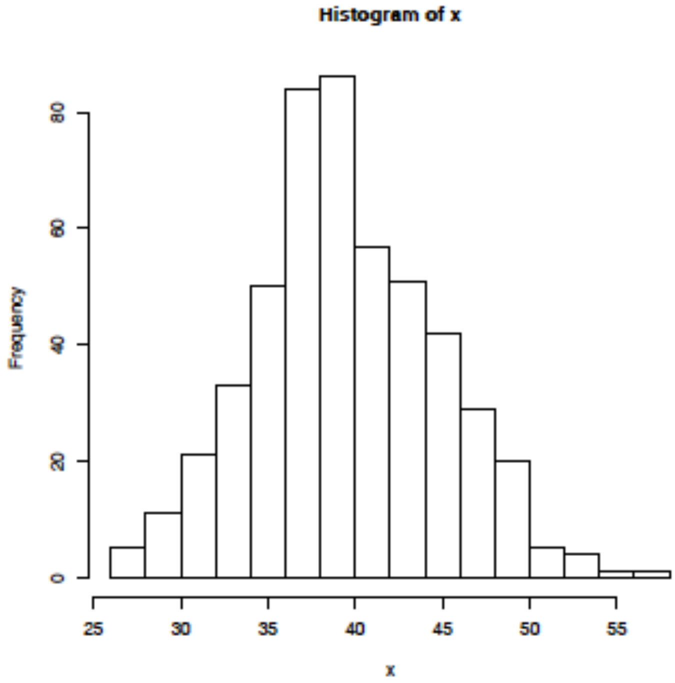
prob. di successo di ogni processo di Bernoulli

$P(X=3)$ $\quad x=3$

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

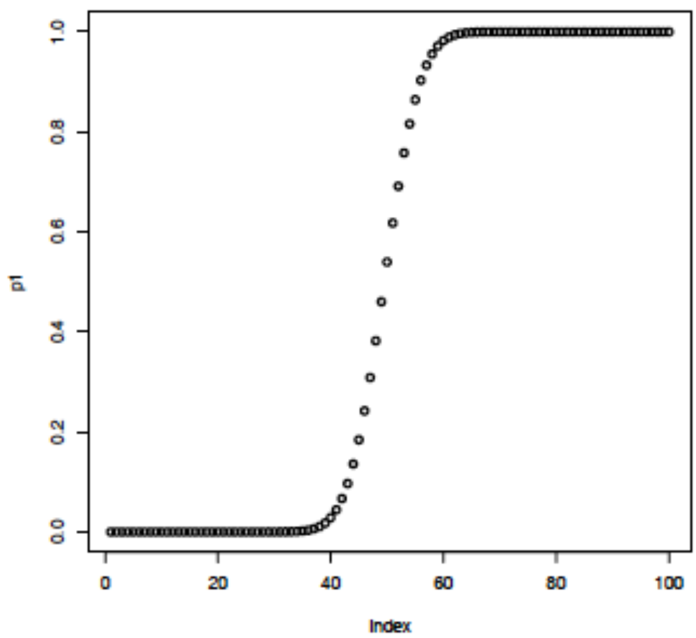
$$f_X(3) = \binom{5}{3} \frac{1}{6}^3 \cdot \frac{5}{6}^2$$

```
>  x <- rbinom(500, size = 100, prob = 0.4)
> hist(x, breaks = 20)
```



Invece per disegnare la CDF:

```
> p1 <- pbinom(1:100, size = 100, prob = 1/2)
> plot(p1)
```



5.2.5 Distribuzione geometrica

La distribuzione geometrica riguarda eventi che dipendono dal tempo. In particolare, se consideriamo una sequenza di processi di Bernoulli e la variabile aleatoria X ={numero di fallimenti prima che ci sia un successo}, questa variabile ha una *distribuzione geometrica di probabilità*. La sua PMF è:

$$f_X(x) = p(1-p)^x, \quad x = 0, 1, 2, \ldots$$

dove p è la probabilità di successo di un singolo processo di Bernoulli. Si indica $X \simeq$ `geom(prob=p)`. Si ha che:

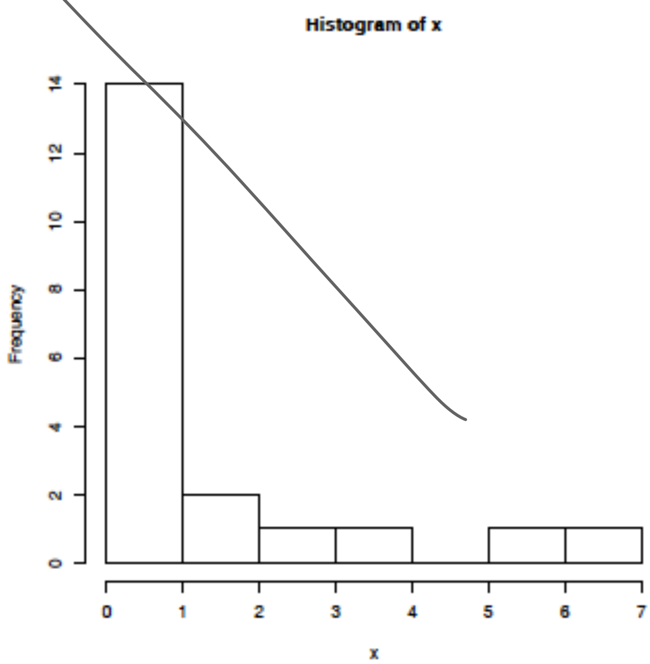
$$\mu = \frac{1-p}{p} \quad \sigma^2 = \frac{1-p}{p^2}.$$

Le funzioni R del pacchetto `prob`, Kerns [2013], sono `dgeom`, `pgeom`, `qgeom`, e `rgeom`.

Esempio 5.26 *Un calciatore ha realizzato il* 90% *dei calci di rigore fino al* 2015. *Qual è la probabilità che sbagli 3 rigori prima di fare un goal?*

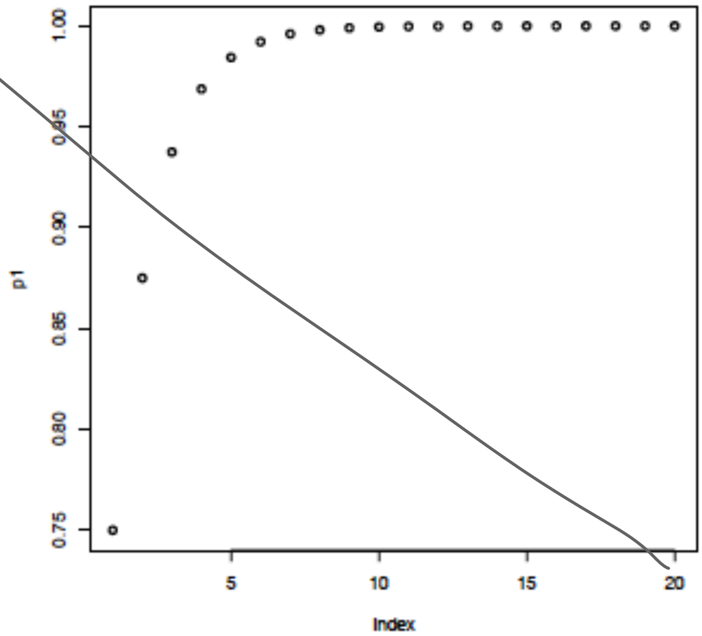
Soluzione. Sia X la variabile aleatoria , X ={numero di rigori sbagliati prima di un successo}, allora $X \simeq$ `geom(prob=0.9)`. Devo quindi considerare $P(X \geq 3) = P(X > 2)$

```
> x <- rgeom(1:20, prob = 1/2)
> hist(x)
```



CDF:

```
> p1 <- pgeom(1:20, prob = 1/2)
> plot(p1)
```




```
> pgeom(2, prob = 0.9, lower.tail = FALSE)

## [1] 0.001
```

5.2.6 Distribuzione di Poisson

La *distribuzione di Poisson* è connessa con eventi 'rari' quali l'arrivo di un cliente in banca o di un'auto dal benzinaio, errori tipografici, ricezione di particelle su un sensore, incidenti, evv. Se λ è la media dell'evento nell'intervallo di tempo $[0, 1]$, e X={numero di eventi che accadono nell'unità di tempo}, allora la PMF di X è:

$$\boxed{f_X(x) = e^{-\lambda}\frac{\lambda^x}{x!}, \quad x = 0, 1, \ldots}$$

e si indica $X \simeq$ `pois(lambda=`λ`)`.

Se invece la variabile aleatoria X ={numero degli eventi che accadono nell'intervallo $[0, t]$, e λ è sempre la media degli eventi che accadono nell'unità di tempo $[0, 1]$, allora si ha che $X \simeq$ `pois(lambda=`λt`)` e la PMF è:

$$f_X(x) = e^{-\lambda t}\frac{(\lambda t)^x}{x!}, \quad x = 0, 1, \ldots$$

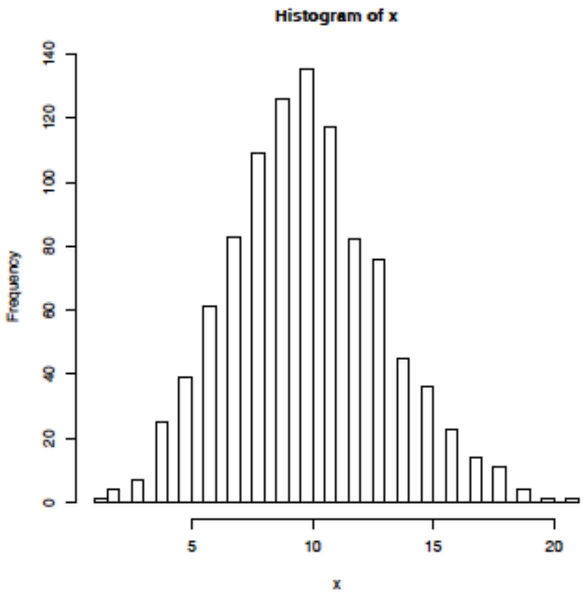
In questo caso media e varianza coincidono:

$$\mu = \sigma^2 = \lambda.$$

In R ci sono le funzioni `dpois`, `ppois`, `qpois`, `rpois`.

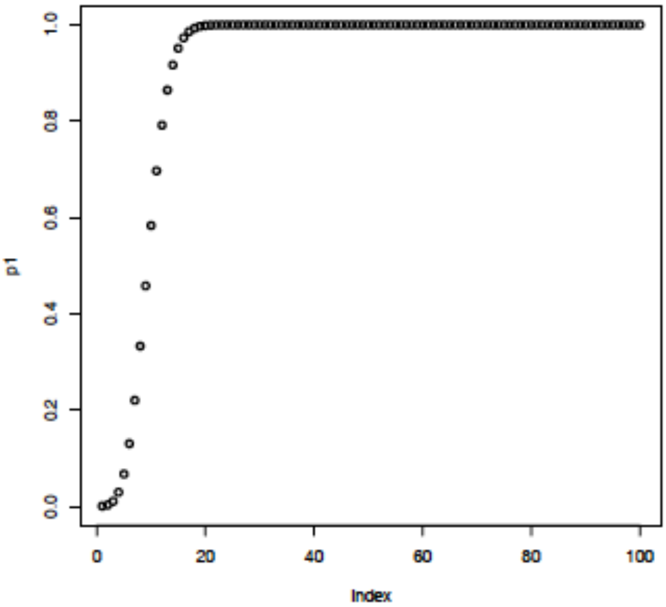
PDF:

```
> x <- rpois(1:1000, lambda = 10)
> hist(x, breaks = 30)
```



CDF:

```
> p1 <- ppois(1:100, lambda = 10)
> plot(p1)
```



$\lambda = 25$

Esempio 5.27 *Supponiamo che dal benzinaio arrivino in media 25 auto all'ora. Sia X la variabile aleatoria X ={numero di auto che arrivano dalle 9 alle 10}.*

1. *Qual è la probabilità che arrivino 10 auto dalle 9 alle 10?*
2. *Qual è la probabilità che fra le 9 e le 10 arrivino fra i 20 e i 30 clienti?*

Soluzione.

1. $X \simeq$ `pois(lambda=25)`.

$$P(X = 10) = f_X(x = 10) = e^{-25}\frac{25^{10}}{10!} = 0.000364....$$

```
> dpois(10, lambda = 25)

## [1] 0.000364985
```

2.
```
> diff(ppois(c(20, 30), lambda = 25))

## [1] 0.6778166
```

$X = \{$ numero di auto che arrivano fra le 9 e le 10 $\}$

25 media delle auto che arrivano in un'ora

X ha distribuzione Poisson $\lambda = 25$

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

$x = 10$

$P(X = 10)$?

$$f_X(10) = e^{-25} \frac{25^{10}}{10!}$$

probabilità che arrivino 10 auto fra le 9 e le 10

$Y = \{$ numero di auto che arrivano fra le 9 e le 11 $\}$

$\lambda_1 = 2 \cdot \lambda$

$P(Y = 10)$? $y = 10$

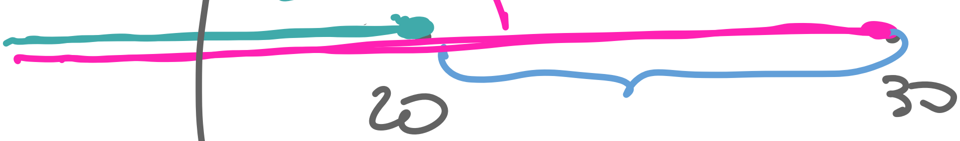
$$f_X(10) = e^{-2 \cdot \lambda} \frac{(2 \cdot \lambda)^y}{y!} =$$

$$= e^{-50} \cdot \frac{50^{10}}{10!}$$

CDF : $F_X(t) = P(X \le t)$

$P(20 \le X \le 30)$?

$P(X \le 30) - P(X \le 20)$



$F_X(30) - F_X(20) = P(20 \le X \le 30)$

ANACONDA

PYTHON

SPYDER

25/2

- Variabile aleatoria
  - discreta
  - continua

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

$S_X \to$ supporto X = l'insieme di tutti i possibili valori di X

DISCRETE

PMF: $f_X : S_X \to [0,1]$

$$f_X(x) = P(X = x)$$

$x \in S_X$

- MEDIA di f_X
- VARIANZA di f_X
- DEVIAZIONE STANDARD di f_X

CDF: $F_X : \mathbb{R} \to [0,1]$

$$F_X(t \in \mathbb{R}) = P(X \leq t)$$

- DISTRIBUZIONE DISCRETA UNIFORME

$$S_x = \{1, \dots n\}$$

$$f_x(x) = \frac{1}{n}$$

S_x

- DISTRIBUZIONE BINOMIALE

$p \to$ prob. di successo di processo Bernoulli

$n \to$ processi di Bernoulli

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

- DISTRIBUZIONE di POISSON

$\lambda \to$ media eventi nell'unità di tempo

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

$$x \in S_x = \{1, 2 \dots\}$$

λ = media di f_x

λ = dev. standard di f_x