distribuzioni teoriche di probabilità

cosa sono

distribuzione teorica di probabilità

- un modello del processo che genera i dati
- specifica la densità di probabilità associata ad ogni possibile valore assunto dai dati
- in molti casi è utile confrontare osservazioni empiriche con un modello

esempio

processo che genera i dati: lanci di un dado onesto

valori possibili:

1, 2, 3, 4, 5, 6

la distribuzione teorica di probabilità è:

1/6, 1/6, 1/6, 1/6, 1/6, 1/6

urna

estrazioni con reimmissione

1, 2, 3, 4, 5, 6

tipologie

discrete: la probabilità si riferisce ad ogni valore di una variabile discreta

continue: la probabilità si riferisce a intervalli di una variabile continua

tipologie

processi diversi sono descritti da distribuzioni teoriche di diversa forma

noi vedremo le distribuzioni: uniforme, binomiale, normale

il dado onesto è una distribuzione uniforme discreta

distribuzione empirica di probabilità

- le frequenze relative osservate per i valori osservati
- quando confrontiamo le osservazioni con il modello, confrontiamo la d.t.p. con la d.e.p.
- ad esempio quando misuriamo l’associazione con chi-quadrato

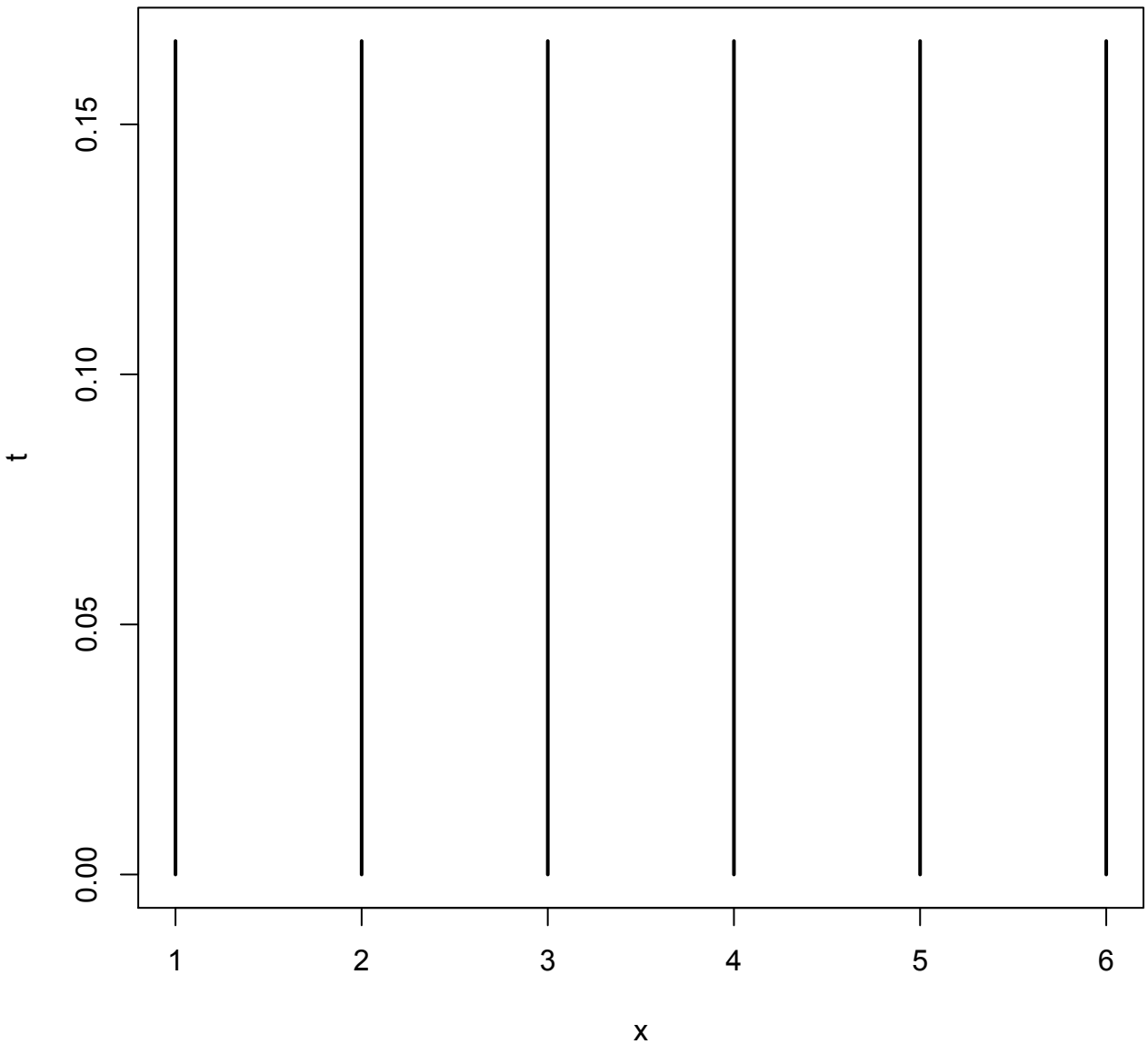
due distribuzioni

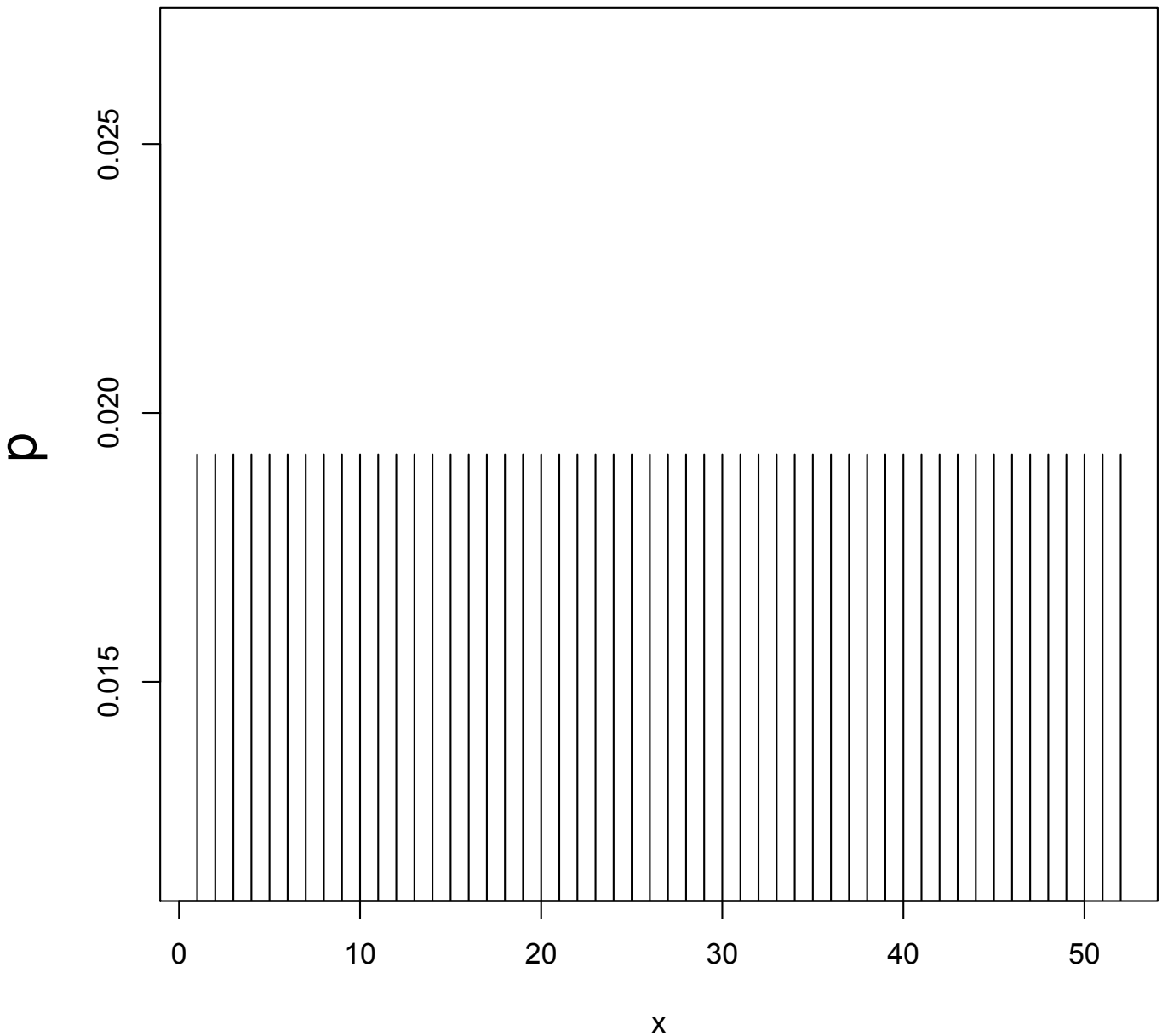
empirica: un vettore di frequenze relative, corrisponde all'informazione raccolta misurando (il campione)

teorica: un vettore di probabilità derivate da un modello (v. discreta) oppure una funzione di densità che descrive il modello(v. continua); corrisponde di solito ad una assunzione sulla popolazione da cui provengono le misure

grafico per la distribuzione teorica di una variabile discreta

```
> x <- 1:6
> t <- table(x)/length(x)
> t
x
      1         2         3
0.1666667 0.1666667 0.1666667

      4         5         6
0.1666667 0.1666667 0.1666667

> plot(t)
```


0.00
0.05
0.10
0.15
t
1
2
3
4
5
6
x


p
0.015
0.020
0.025
0
10
20
30
40
50
x

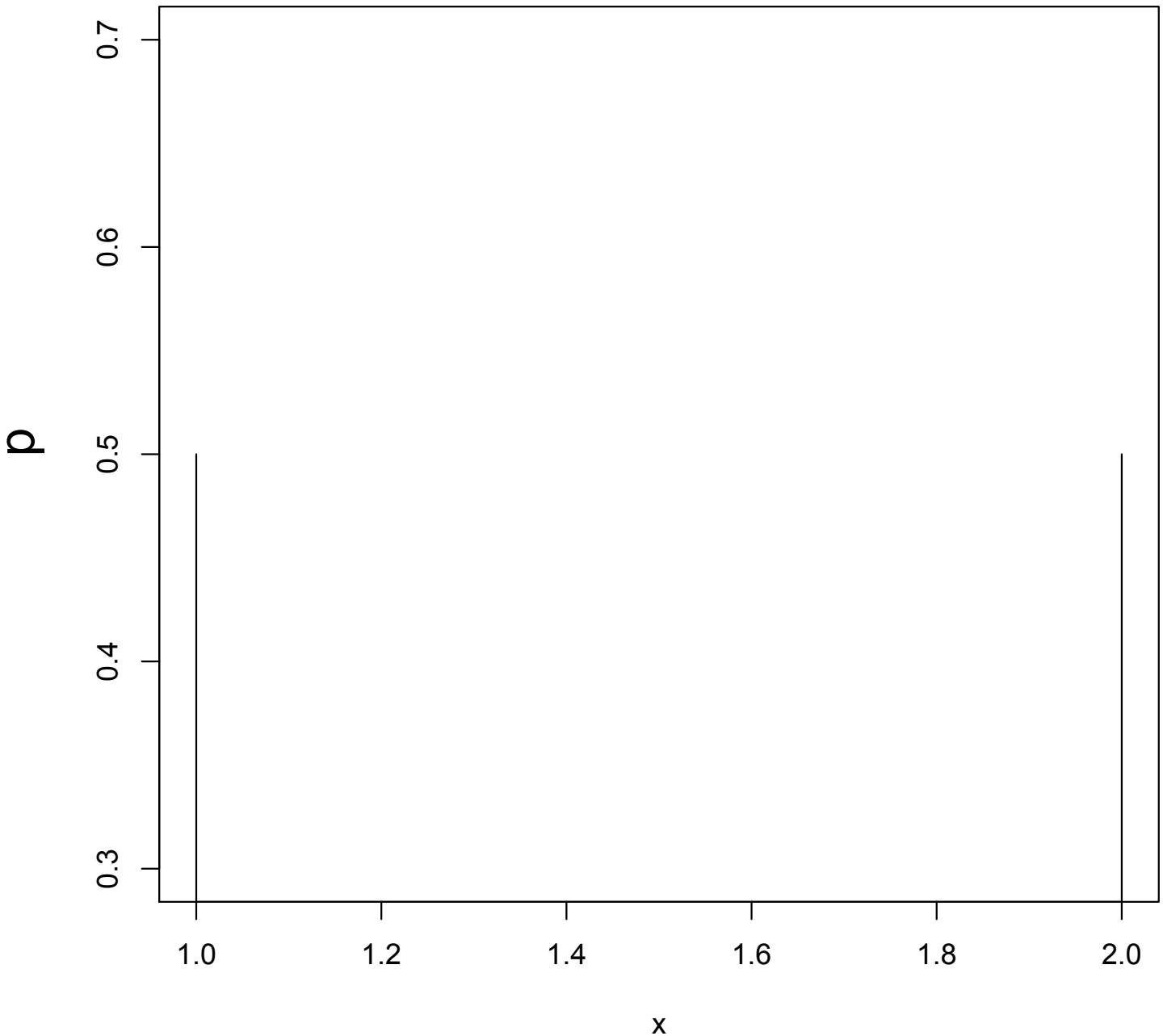
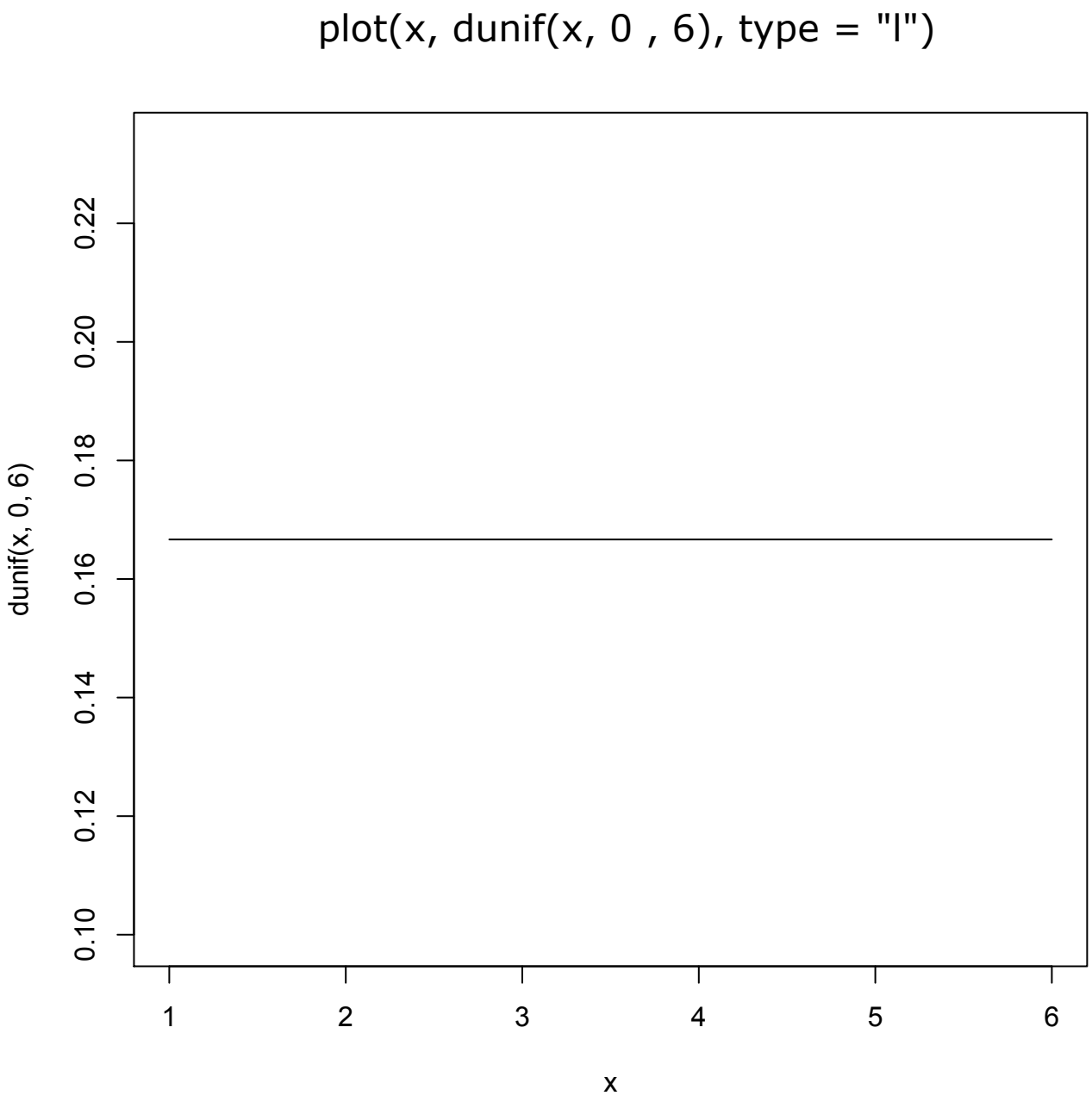

p
0.3
0.4
0.5
0.6
0.7
1.0
1.2
1.4
1.6
1.8
2.0
x

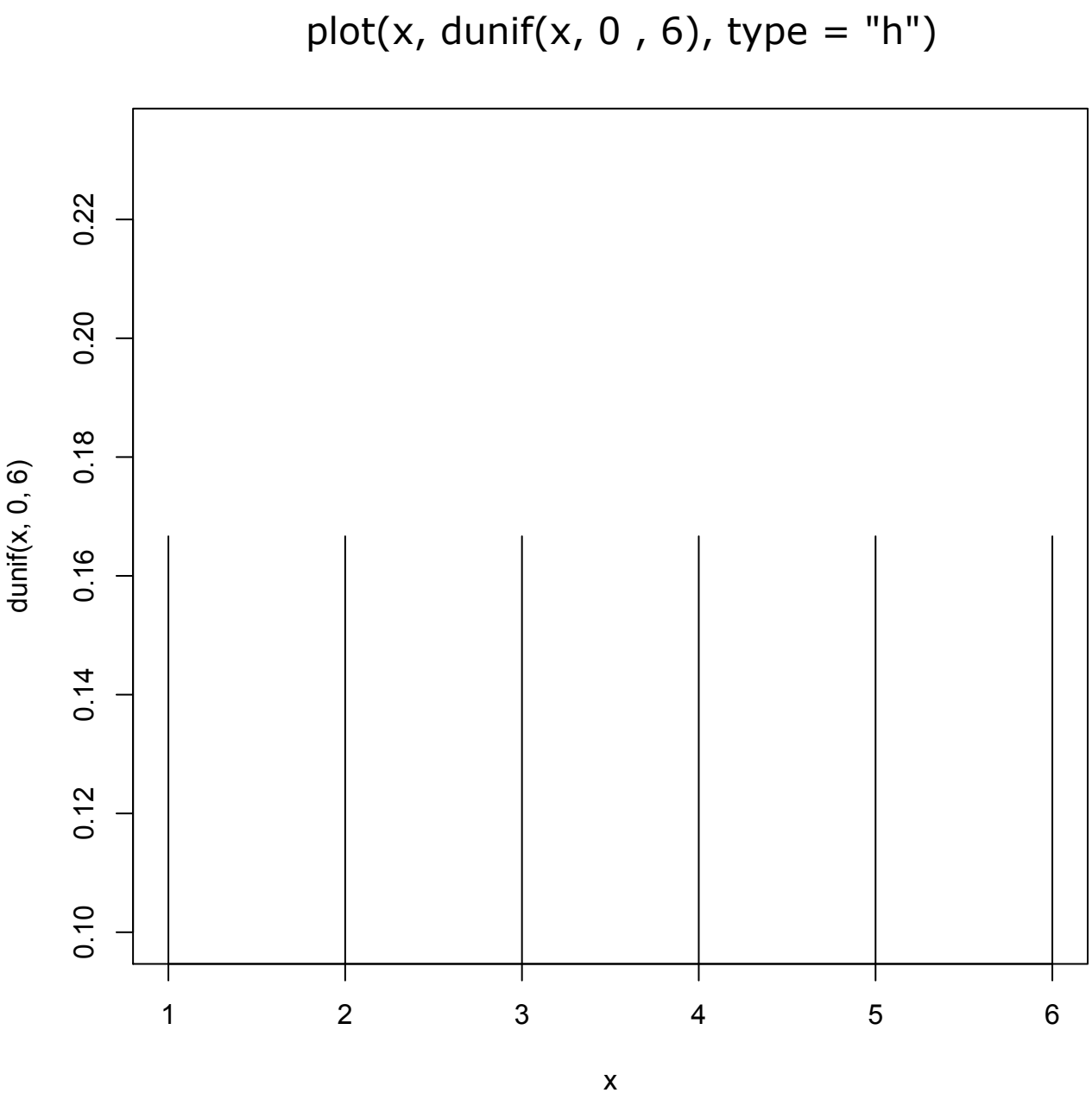
grafico per la distribuzione teorica di una variabile continua

densità della distribuzione uniforme:

$$d = \frac{1}{(MAX - MIN)}$$

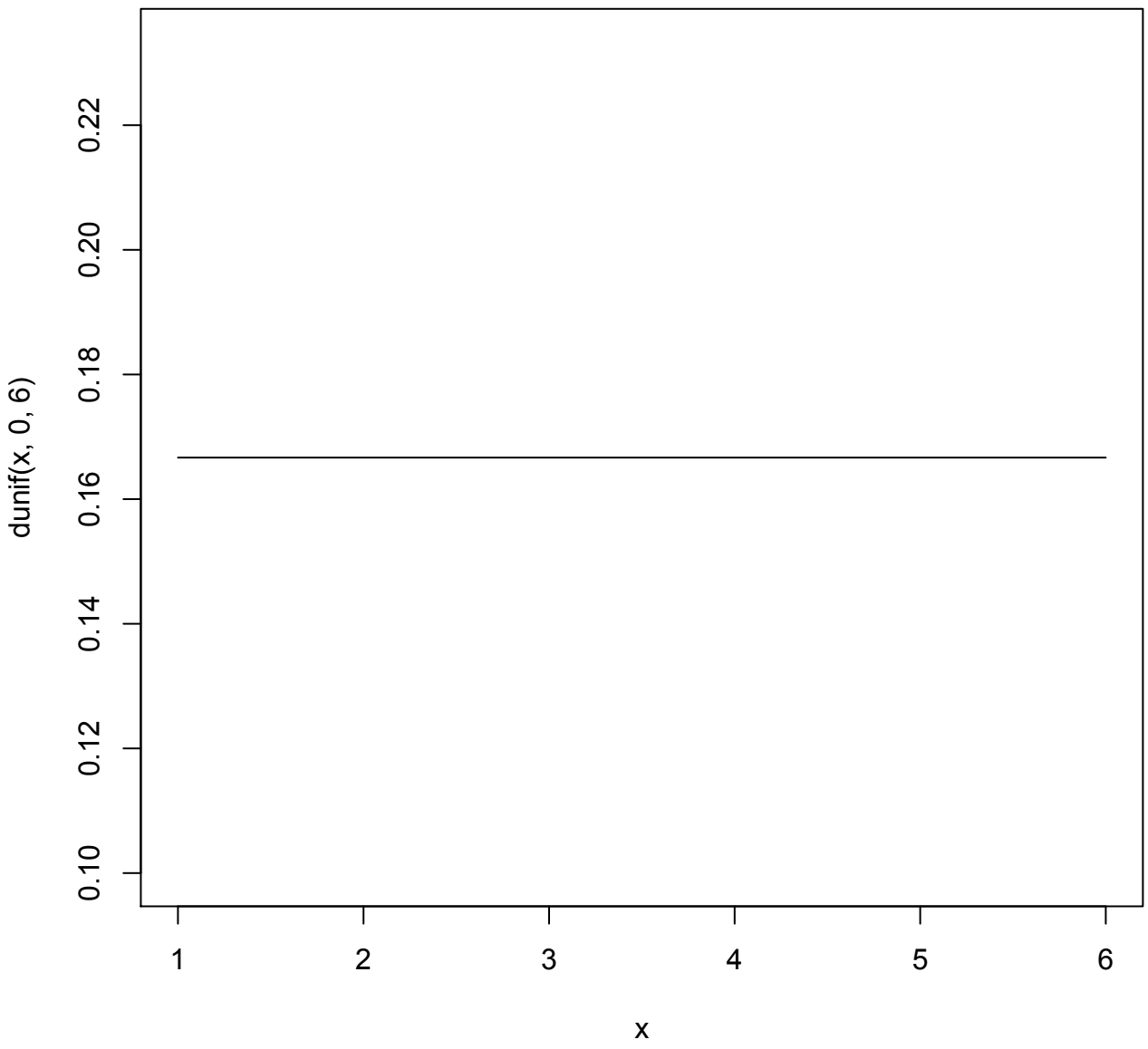
```
> x <- 1:6
> plot(x, dunif(x, 0 , 6), type = "l")
```


plot(x, dunif(x, 0 , 6), type = "l")
dunif(x, 0, 6)
0.10
0.12
0.14
0.16
0.18
0.20
0.22
1
2
3
4
5
6
x


plot(x, dunif(x, 0 , 6), type = "h")
dunif(x, 0, 6)
0.10
0.12
0.14
0.16
0.18
0.20
0.22
1
2
3
4
5
6
x

```
> x <- 1:6
> dunif(x, 1 , 6)
[1] 0.2 0.2 0.2 0.2 0.2 0.2

> x <- 1:2
> dunif(x, 0, 6)
[1] 0.1666667 0.1666667

> dunif(x, 0, 100)
[1] 0.01 0.01

> dunif(x, 10, 100)
[1] 0 0
```

```
> x <- seq(0, 6, by = 0.5)

> x
 [1] 0.0 0.5 1.0 1.5 2.0 2.5 3.0 3.5 4.0 4.5 5.0
5.5 6.0

> dunif(x, 0, 6)
 [1] 0.1666667 0.1666667 0.1666667
0.1666667 0.1666667 0.1666667
 [7] 0.1666667 0.1666667 0.1666667
0.1666667 0.1666667 0.1666667
[13] 0.1666667

> x <- seq(0, 6, by = 0.01)
> plot(x, dunif(x, 0, 6), type = "l")
```


dunif(x, 0, 6)
0.10
0.12
0.14
0.16
0.18
0.20
0.22
1
2
3
4
5
6
x

distribuzione di probabilità cumulativa

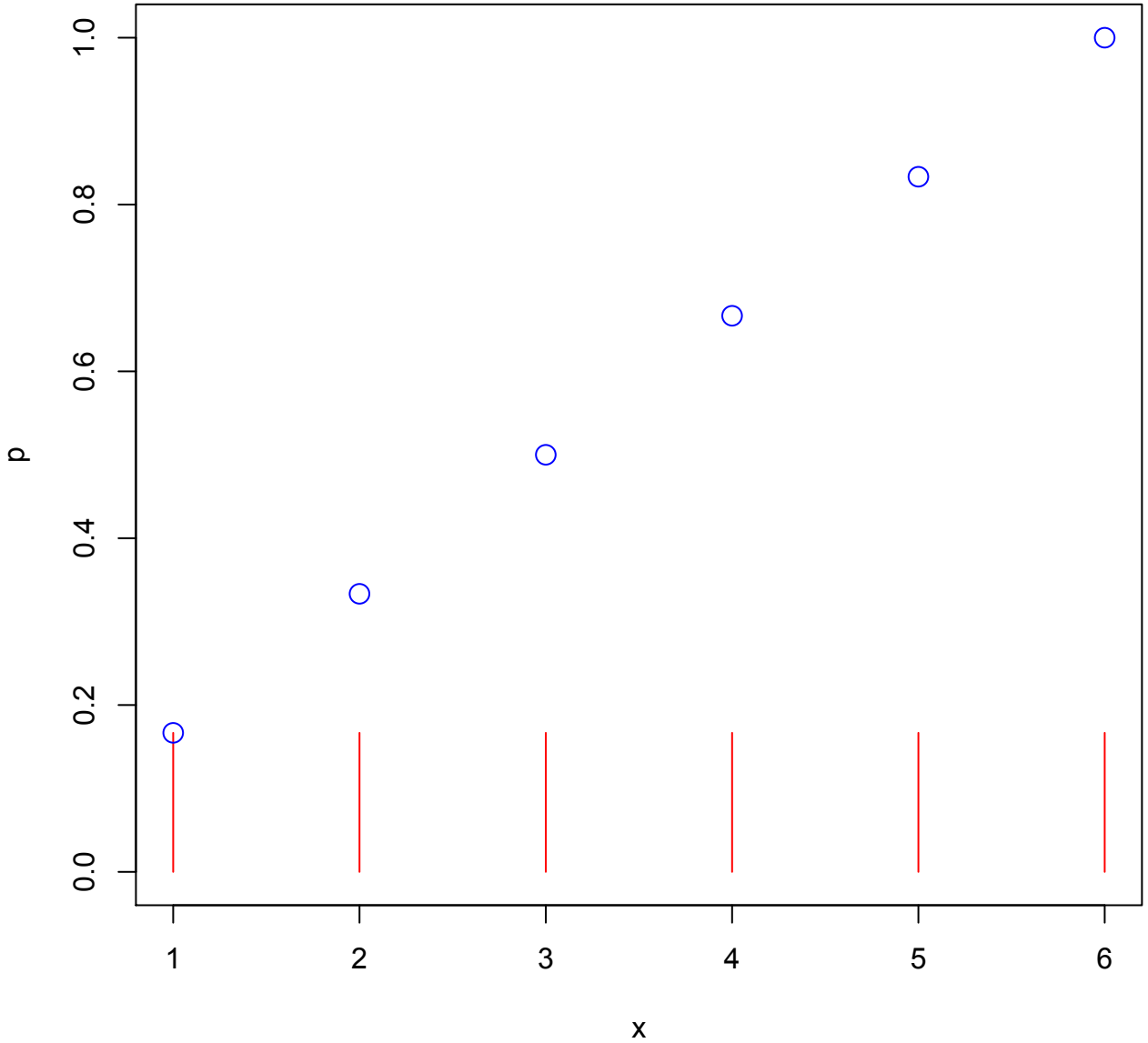
1, 2, 3, 4, 5, 6

la distribuzione teorica di probabilità è:

1/6, 1/6, 1/6, 1/6, 1/6, 1/6

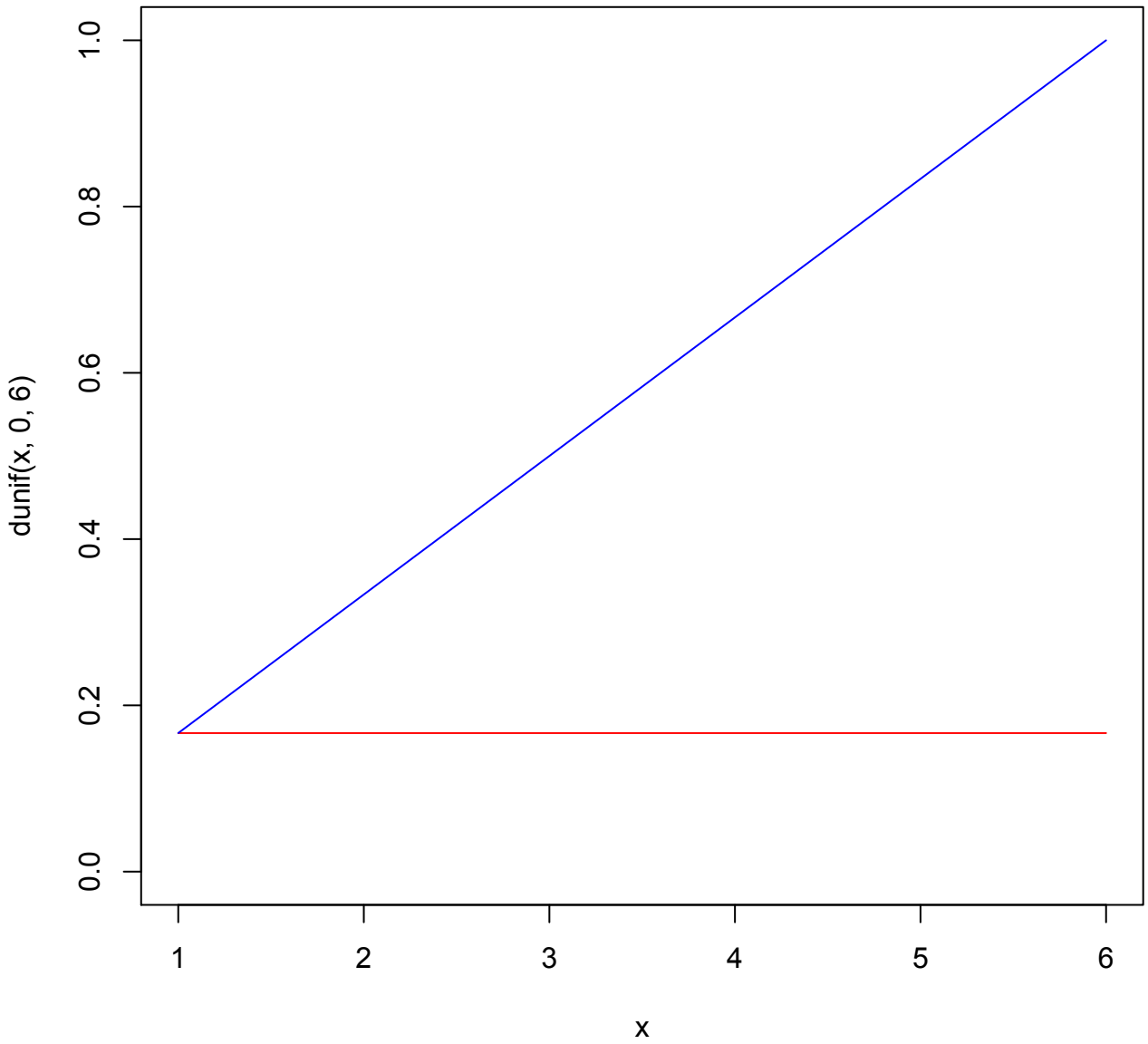
la distribuzione di probabilità cumulativa è:

1/6, 2/6, 3/6, 4/6, 5/6, 6/6



probabilità cumulativa (discreta)

probabilità



probabilità cumulativa

densità

lavorare con la distribuzione uniforme (continua)

dunif()

- densità

punif()

- probabilità cumulativa da MIN a x

qunif()

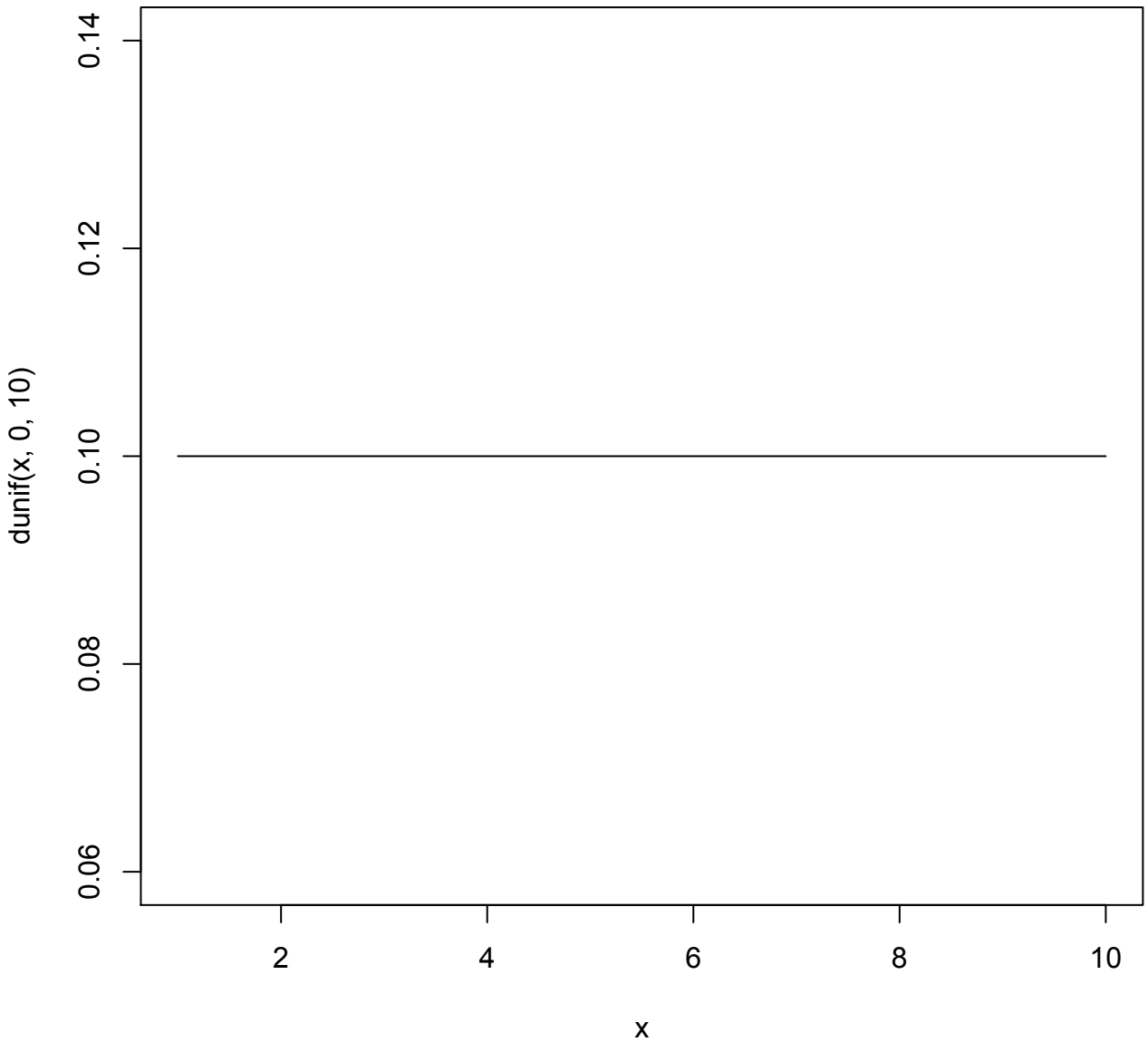
- inversa di punif()

runif()

- campione casuale di grandezza n

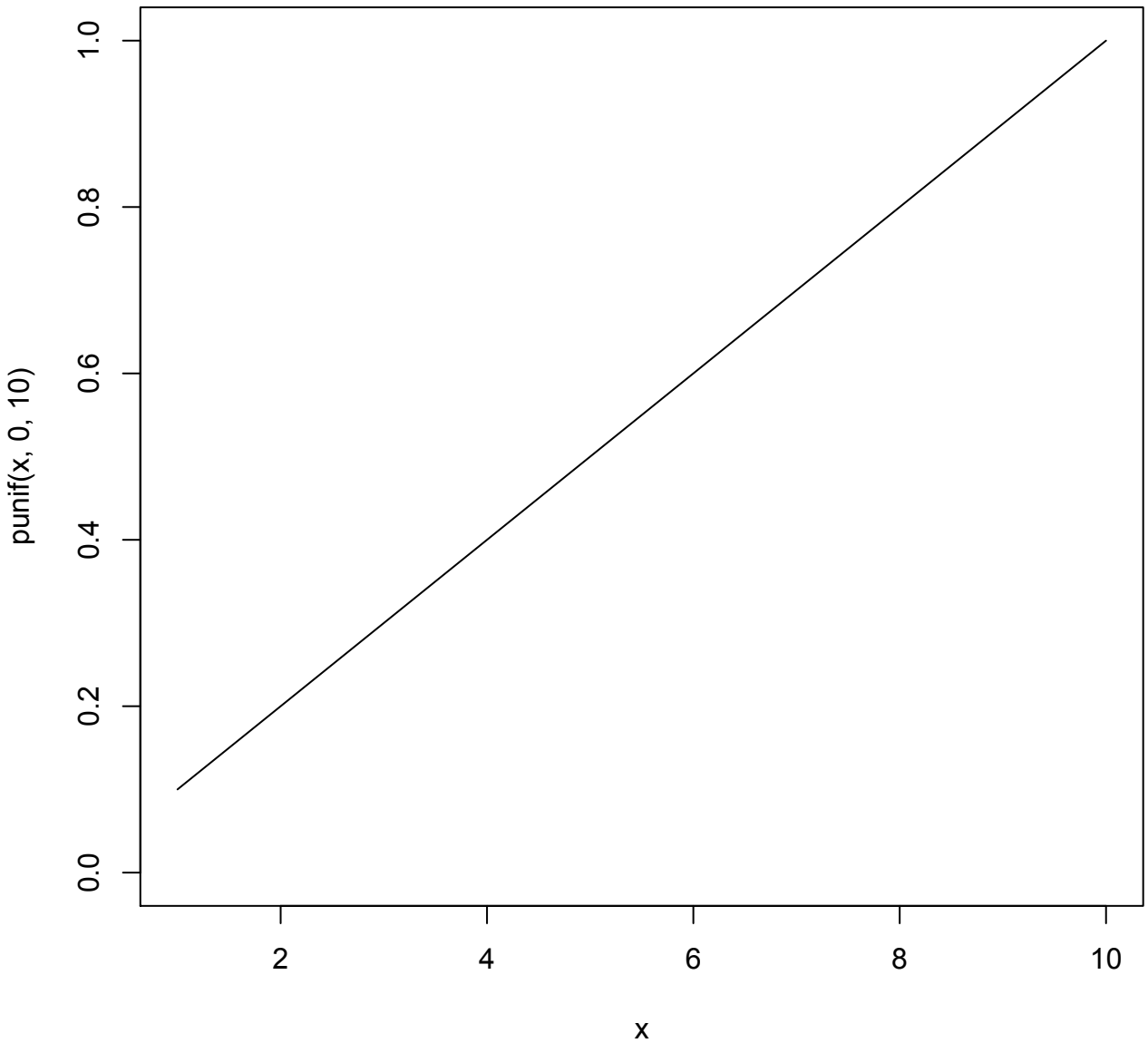
```
> x <- 1:10

> dunif(x, 0, 10)

 [1] 0.1 0.1 0.1 0.1 0.1 0.1 0.1 0.1 0.1 0.1
```

```
> plot(x, dunif(x, 0 ,10), type = "l")
```



```
> x <- 1:10

> punif(x, 0, 10)
 [1] 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0
```

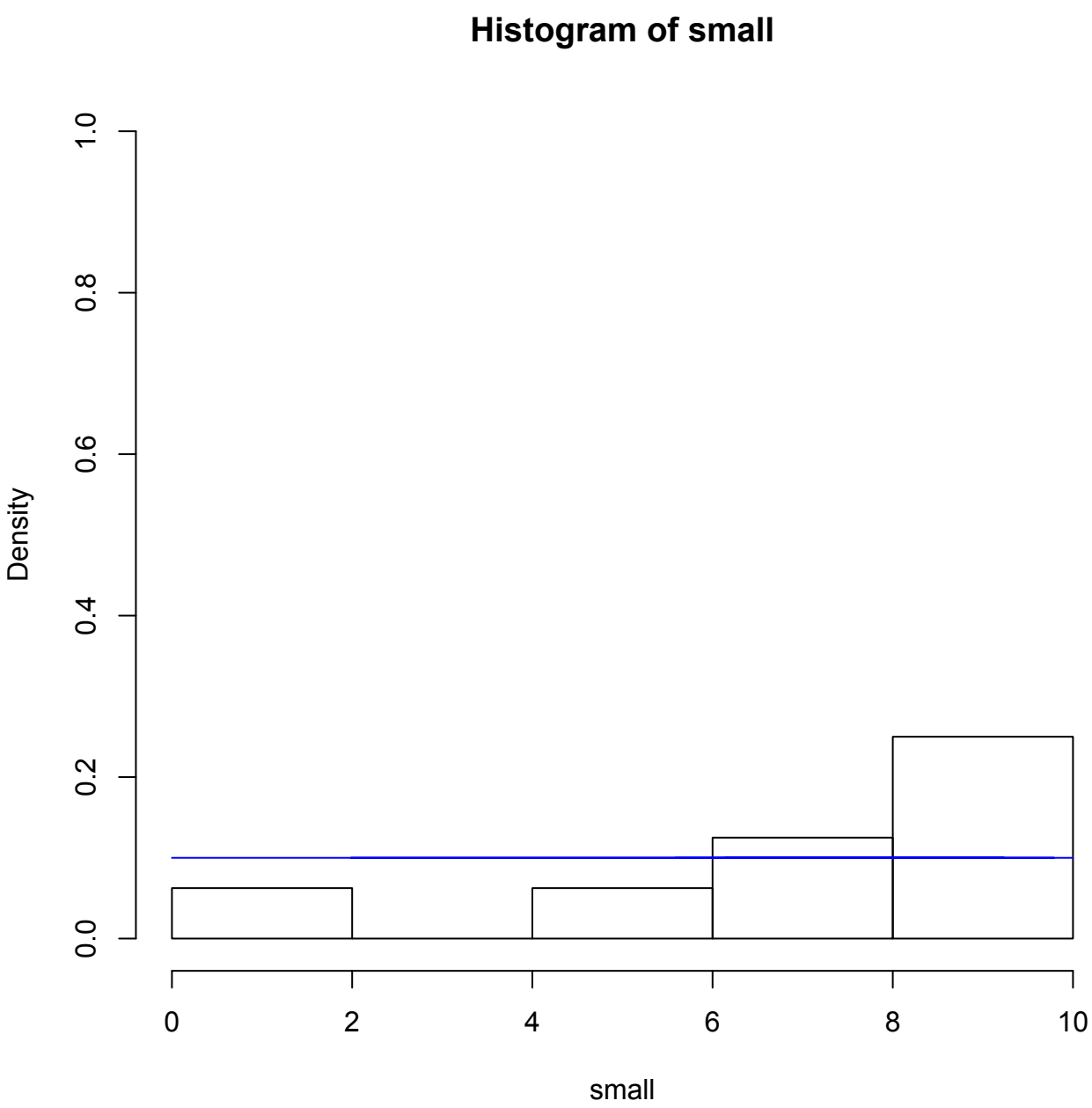
```
> plot(x, punif(x, 0, 10), type = "l", ylim = c(0, 1))
```



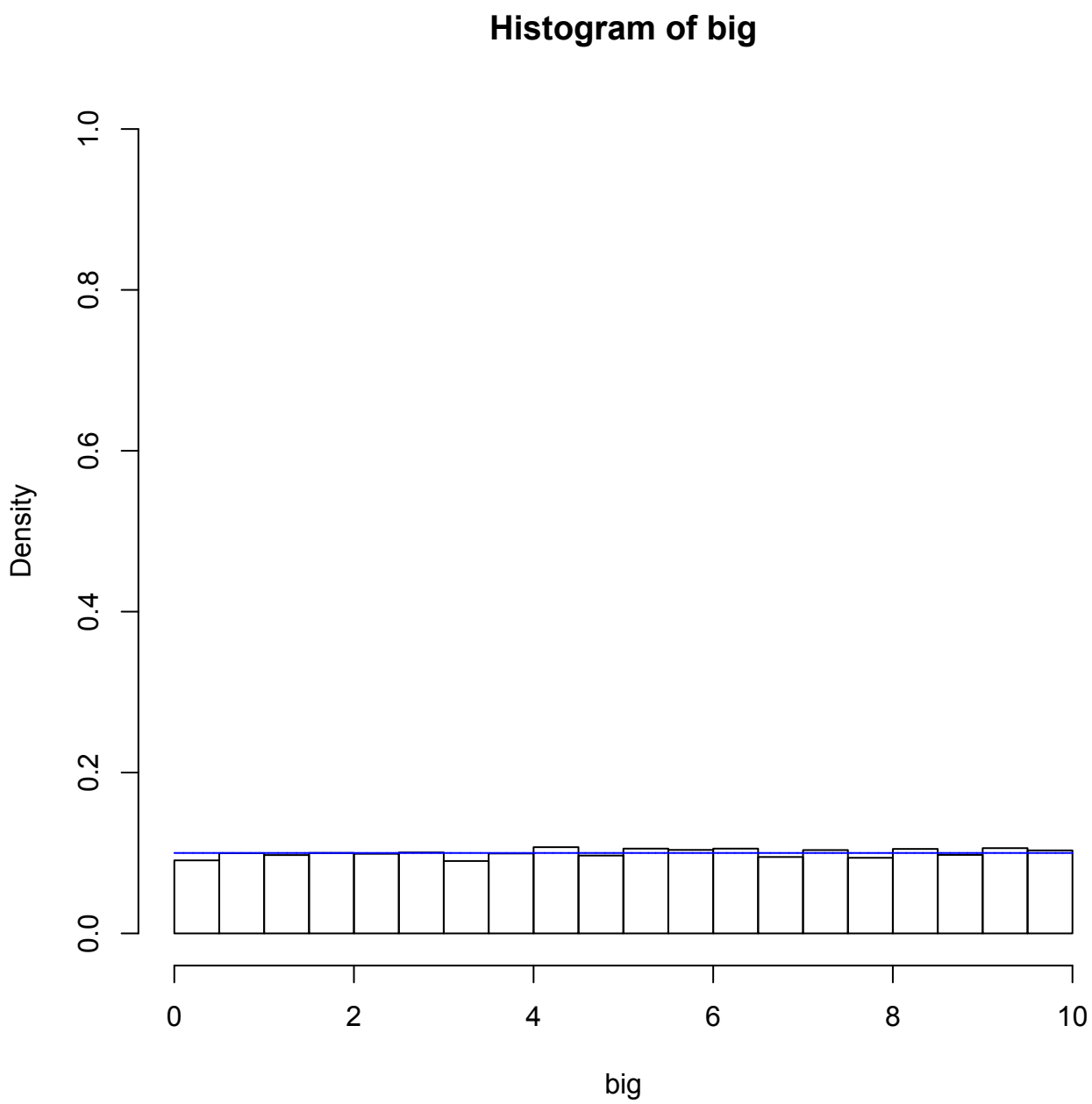
```
> x <- 1:10

> x
 [1]  1  2  3  4  5  6  7  8  9 10

> qunif(0.1, 0, 10)
[1] 1

> qunif(0.9, 0, 10)
[1] 9

> qunif(0.5, 0, 10)
[1] 5
```

```
> x <- 1:10

> x
 [1]  1  2  3  4  5  6  7  8  9 10

> runif(3, 0, 10)
[1] 3.9965445 6.2899339 0.3376736

> runif(10, 0, 10)
 [1] 0.1362454 2.6288281 8.1982060 4.2602240
   0.4094486 2.6150192
 [7] 7.9956288 8.5632740 1.0989128 2.3210183
```

```
> small <- runif(8, 0, 10)

> hist(small, prob = "TRUE", ylim = c(0,
  1))



> lines(0:10, dunif(0:10, 0, 10), col =
  "blue")
```


Histogram of small
Density
0.0
0.2
0.4
0.6
0.8
1.0
0
2
4
6
8
10
small

```
> big <- runif(10000, 0, 10)

> hist(big, prob = "TRUE", ylim = c(0,
  1))

> lines(big, dunif(big), 0, 10), col =
  "blue")
```


Histogram of big
Density
0.0
0.2
0.4
0.6
0.8
1.0
0
2
4
6
8
10
big

la distribuzione normale o gaussiana

densità normale

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

π = pi greco (3.14....)
e = numero di Nepero (2.718....)

μ = media
σ = deviazione standard

normale standard

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

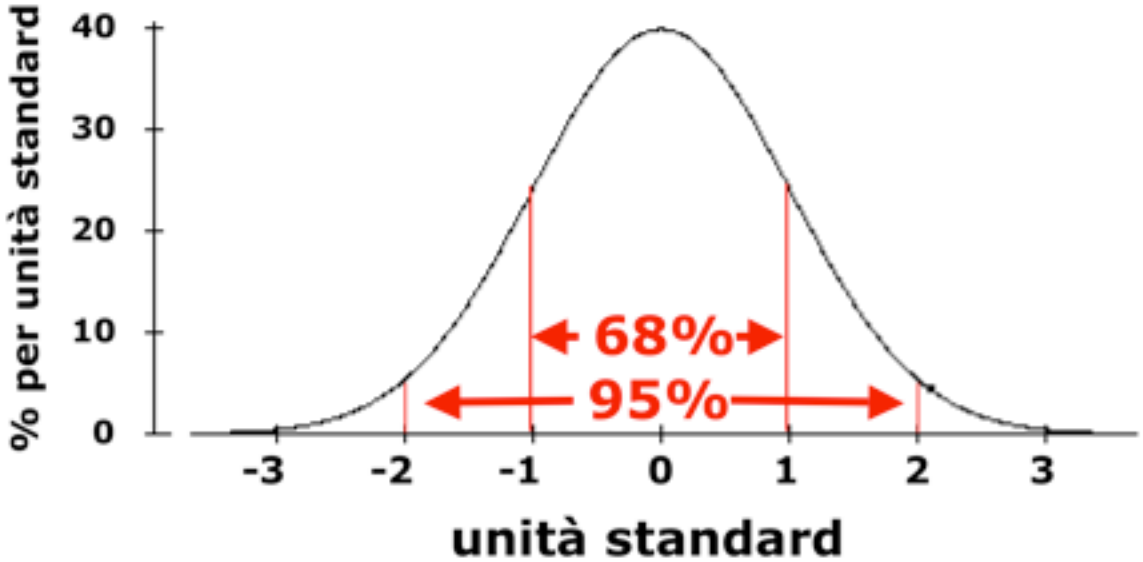
$$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}.$$

$\mu = 0$
$\sigma = 1$

densità normale

“campana” simmetrica e unimodale
media = mediana = moda
x da -INF a +INF (asintotica)
y (densità) da 0 a 0.3989
MAX = media
punti di flesso = -1 DS e + 1 DS
area da -INF a X = prob. cumulativa
area fra - 1 DS e + 1 DS c.a. 68%
area fra - 2 DS e + 2 DS c.a. 95%
area totale sotto la curva = 100%

forma e proprietà



DS

da dove viene

Abraham de Moivre
(1667-1754)

Carl Friedrich Gauss
(1777-1855)

legge dell’errore

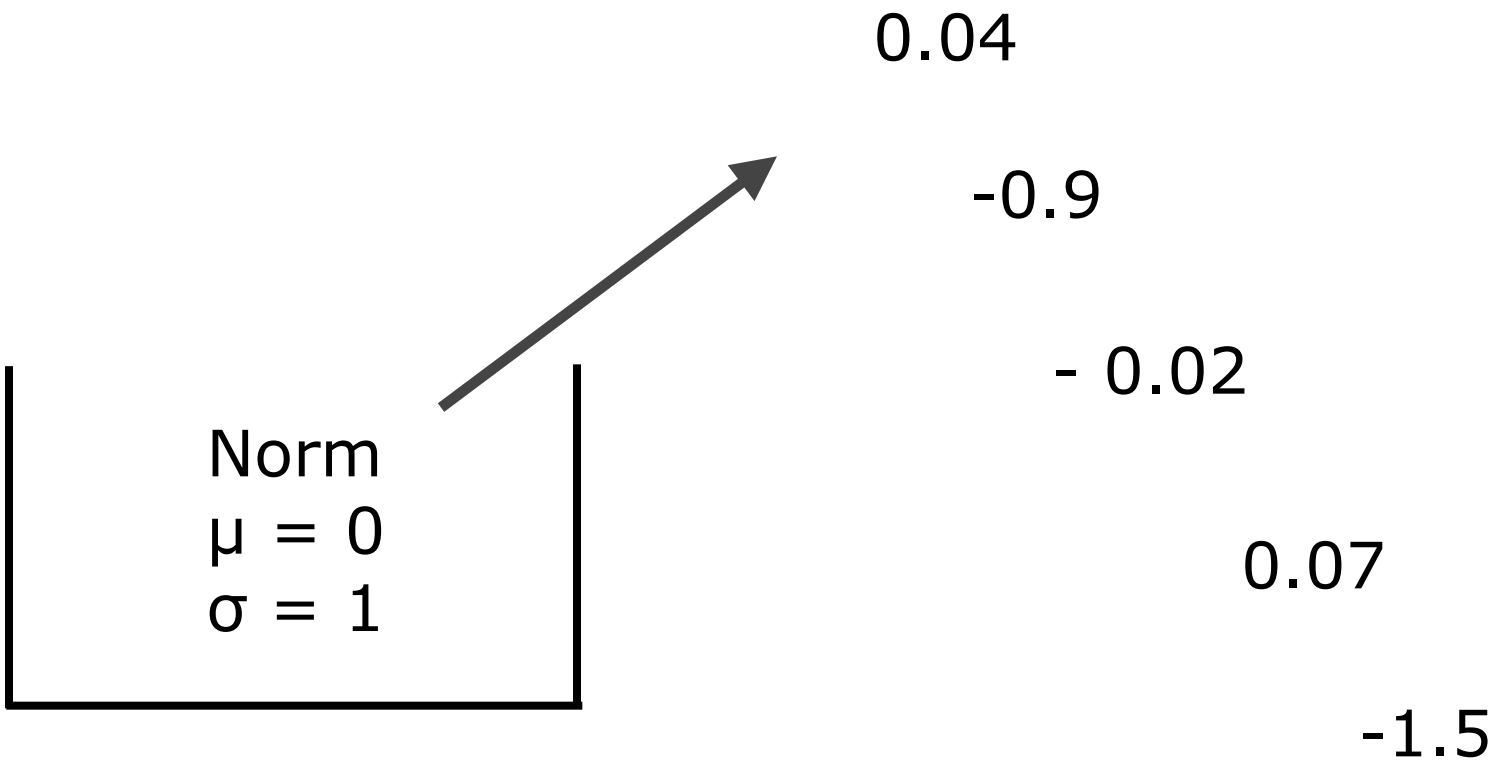
la curva normale è un modello dell’errore casuale

gli errori più probabili stanno attorno alla media (0)

allontanandosi dalla media, gli errori diventano sempre più grandi, ma sempre più improbabili

gli errori casuali sono simmetrici

urna



lavorare con la distribuzione normale

dnorm()

densità

pnorm()

probabilità cumulativa da -INF a x

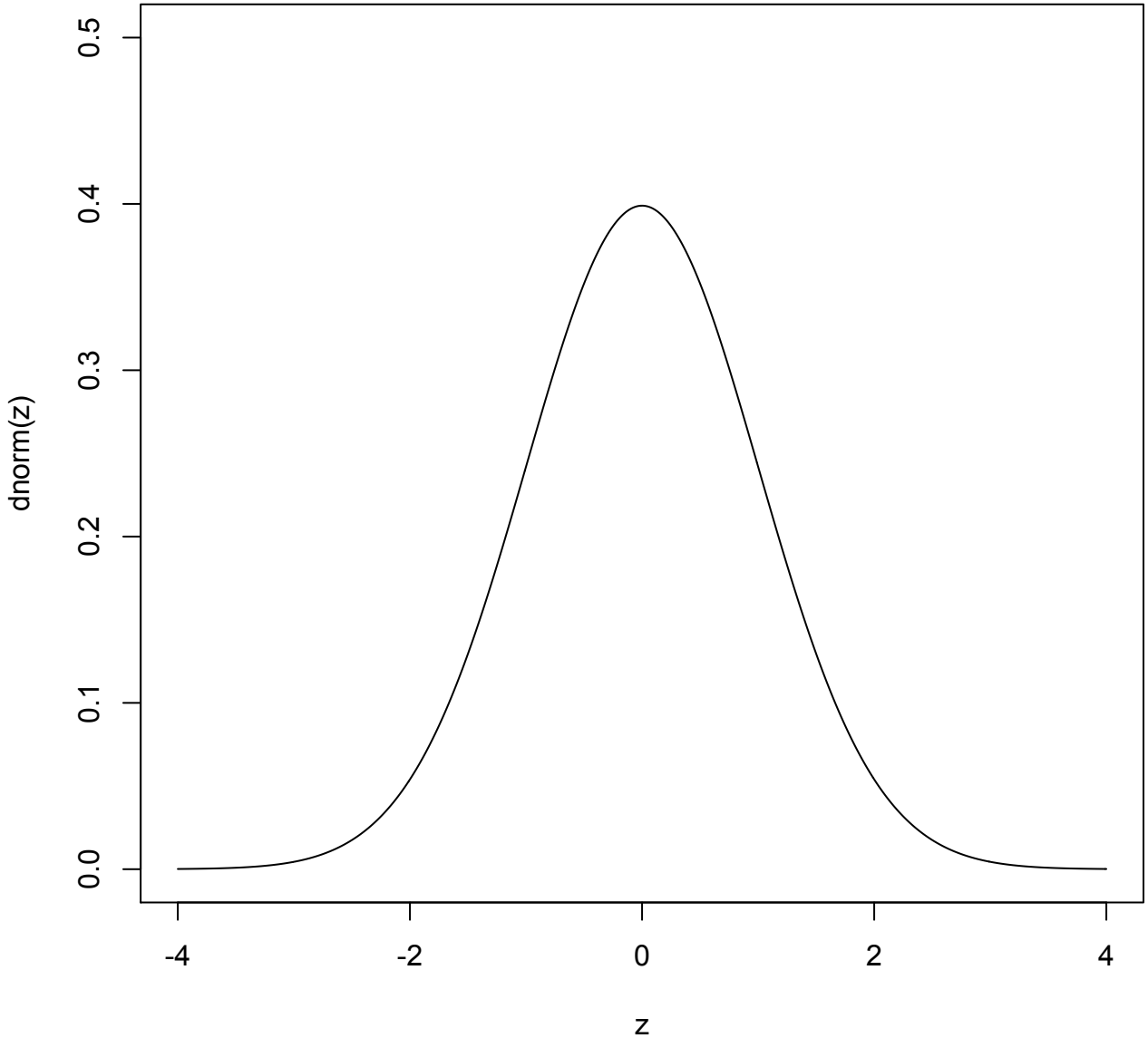
qnorm()

inversa di pnorm()

rnorm()

campione casuale di grandezza n

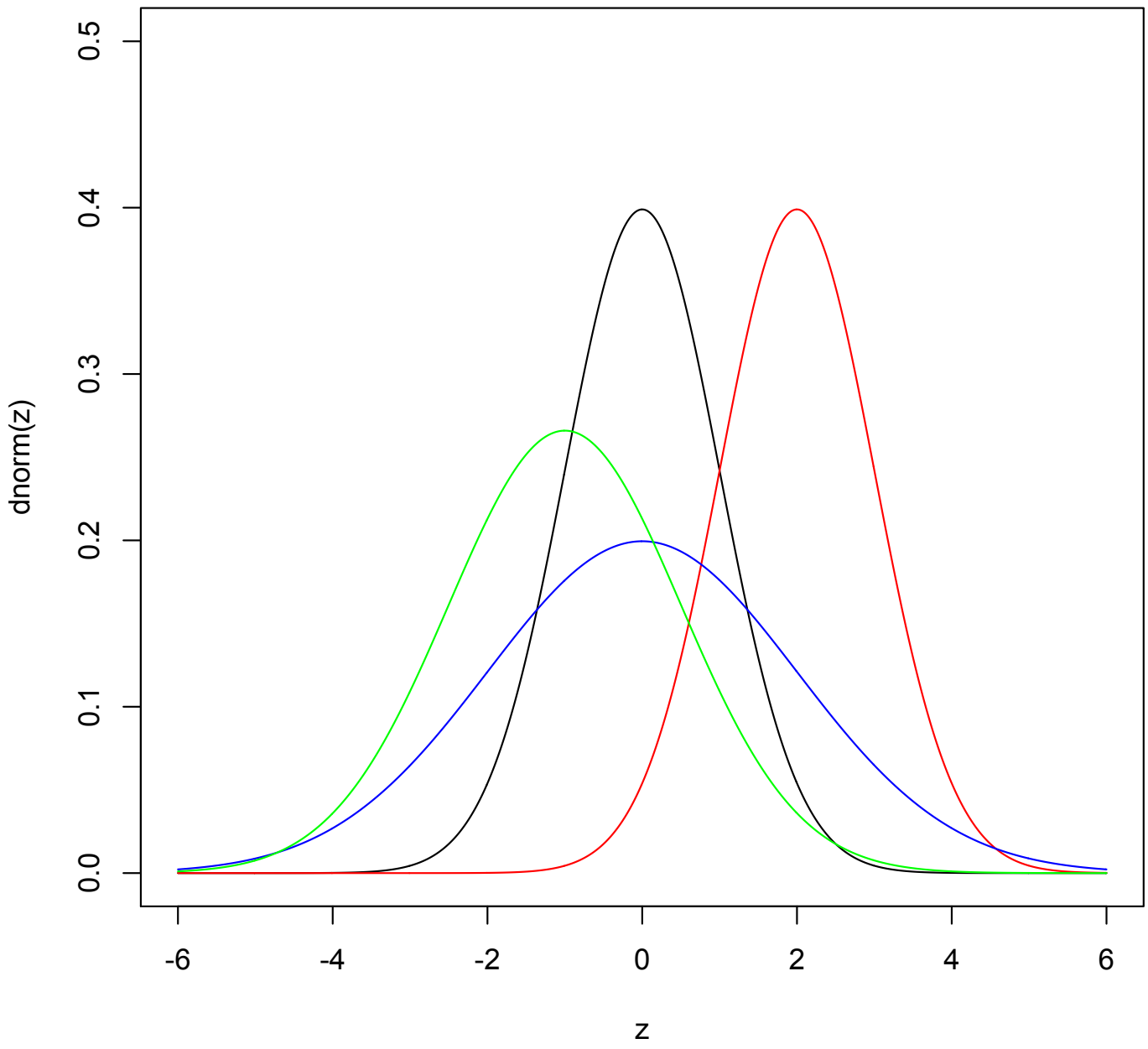
```
> z <- seq(-4, 4, by = 0.01)
> plot(z, dnorm(z), type = "l", ylim = c(0, 0.5))
```



```
> dnorm(0)
[1] 0.3989423
> dnorm(-1)
[1] 0.2419707
> dnorm(1)
[1] 0.2419707
> dnorm(2)
[1] 0.05399097
> dnorm(3)
[1] 0.004431848
> dnorm(4)
[1] 0.0001338302
```

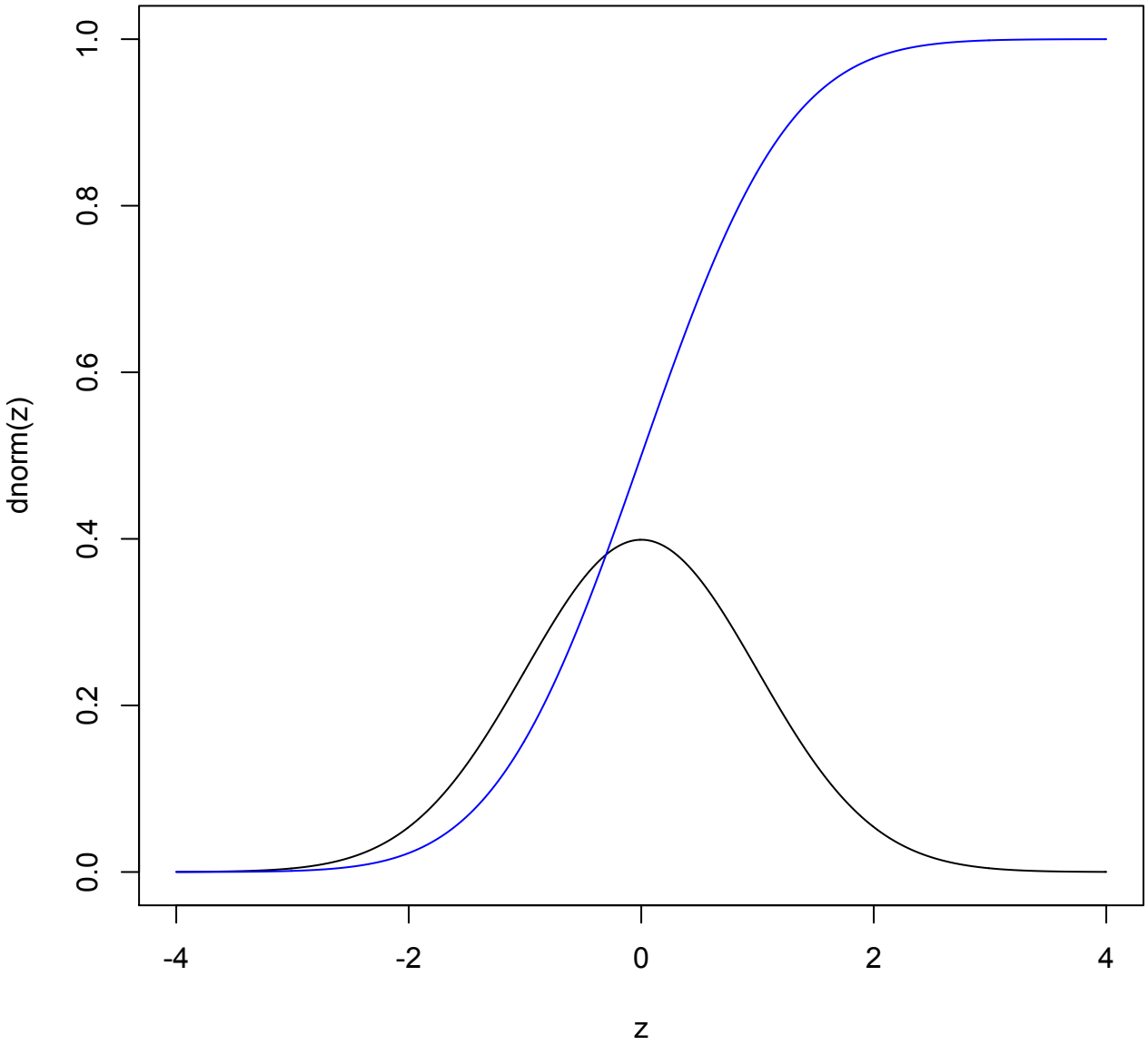
```
> dnorm(0)
[1] 0.3989423

> dnorm(0, mean = 3)
[1] 0.004431848

> dnorm(0, sd = 10)
[1] 0.03989423

> dnorm(0, mean = 3, sd = 10)
[1] 0.03813878
```

```
> z <- seq(-6, 6, by = 0.01)

> plot(z, dnorm(z), type = "l", ylim = c(0,
  0.5))
> lines(z, dnorm(z, mean = 2), col =
  "red")
> lines(z, dnorm(z, sd = 2), col = "blue")
> lines(z, dnorm(z, mean = -1, sd = 1.5),
  col = "green")
```


0.5
0.4
0.3
0.2
0.1
0.0
dnorm(z)
-6
-4
-2
0
2
4
6
z

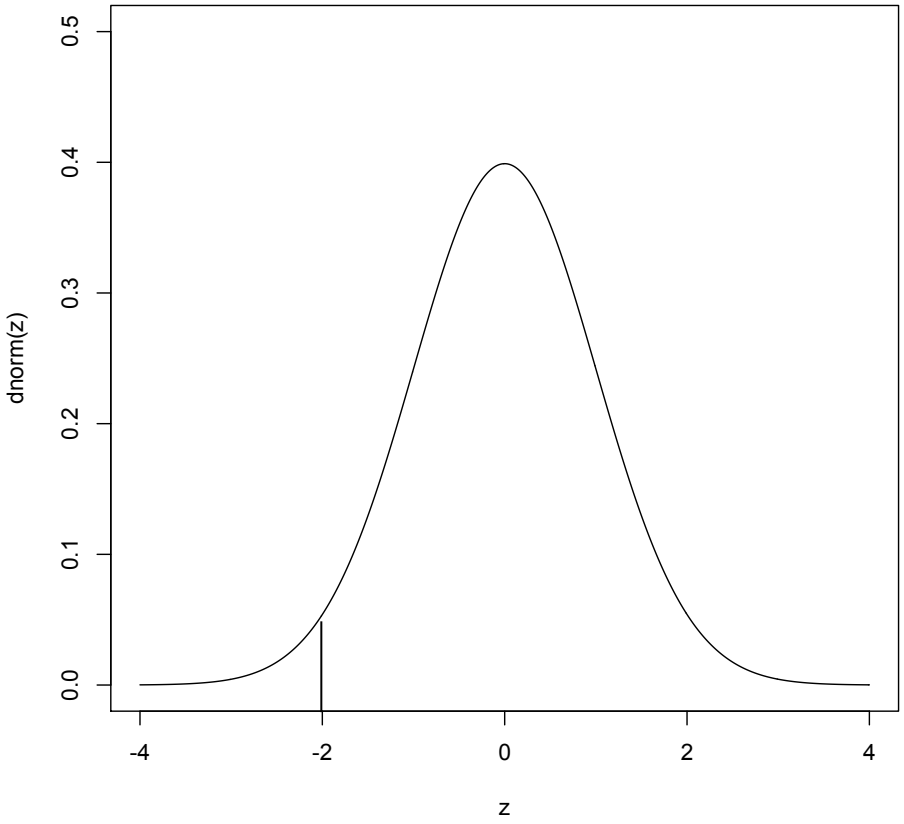
```
> pnorm(0)
[1] 0.5
> pnorm(1)
[1] 0.8413447
> pnorm(-1)
[1] 0.1586553
> pnorm(-3)
[1] 0.001349898
> pnorm(3)
[1] 0.9986501
```

```
> z <- seq(-4, 4, by = 0.01)

> plot(z, dnorm(z), type = "l",
                        ylim = c(0, 1))

> lines(z, pnorm(z), col = "blue")
```


1.0
0.8
0.6
0.4
0.2
0.0
dnorm(z)
-4
-2
0
2
4
z

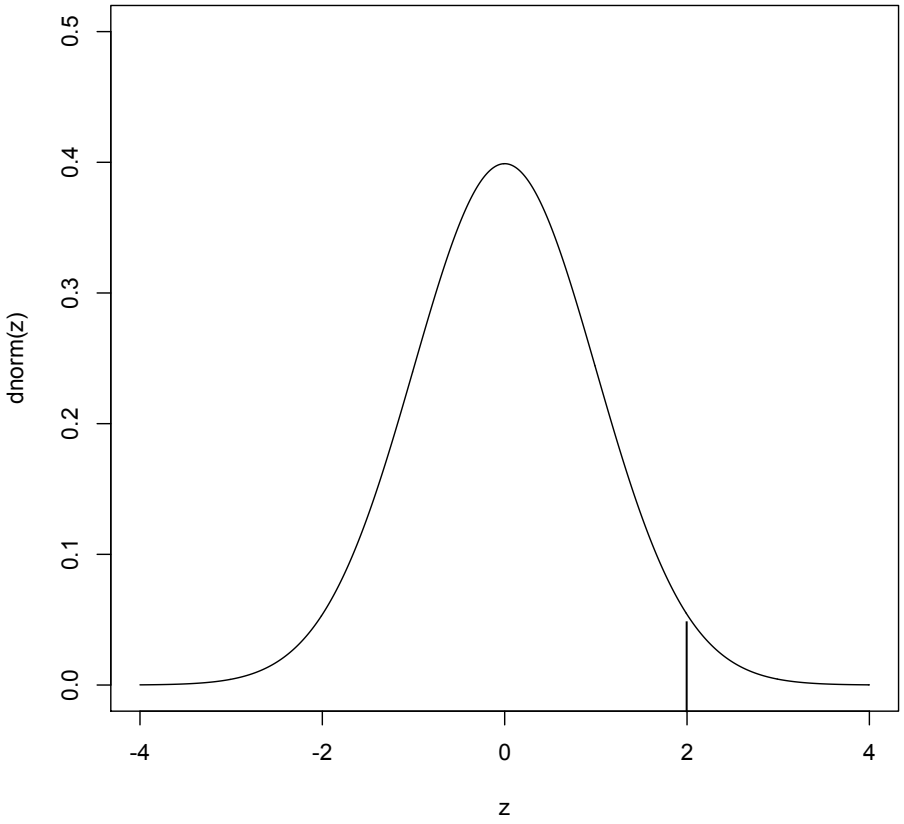
area a sinistra di 2 DS:

```
> pnorm(-2)
[1] 0.02275013
```



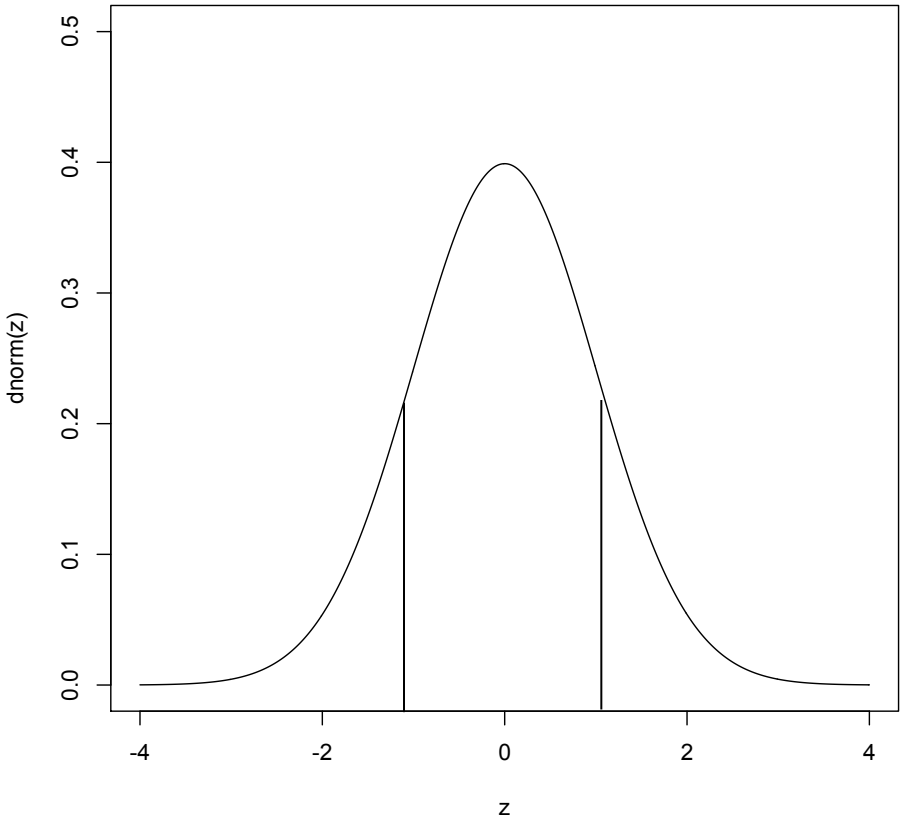
area a destra di 2 DS:

```
> 1 - pnorm(2)
[1] 0.02275013
```



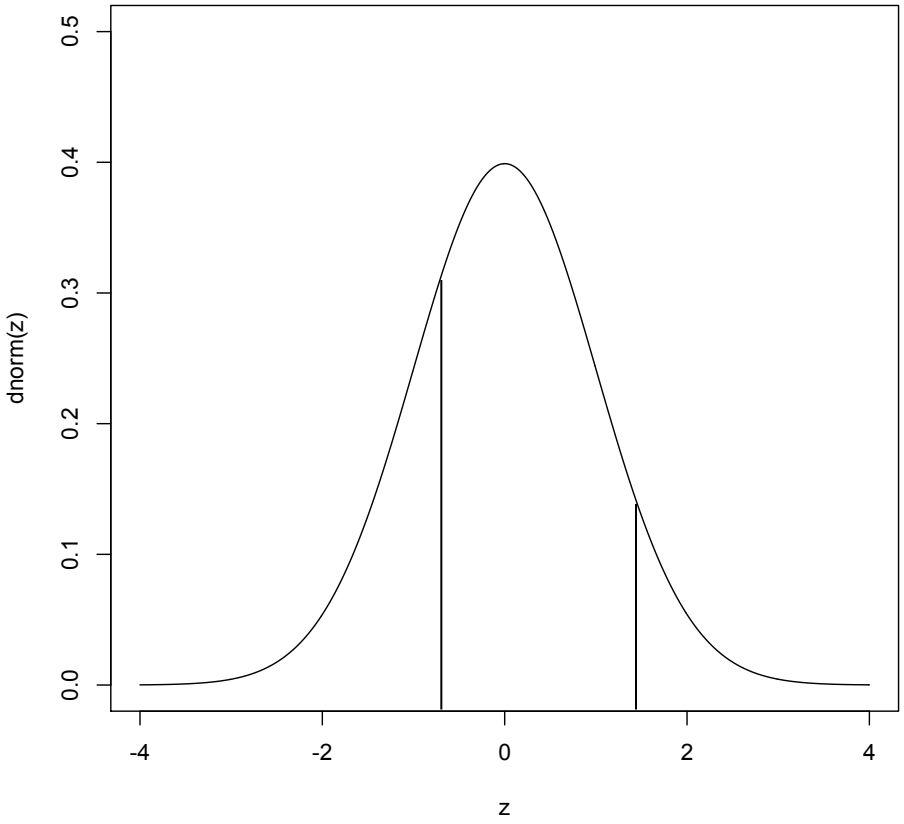
area sotto la curva fra -1 e +1 DS:

```
> pnorm(1) - pnorm(-1)
[1] 0.6826895
```



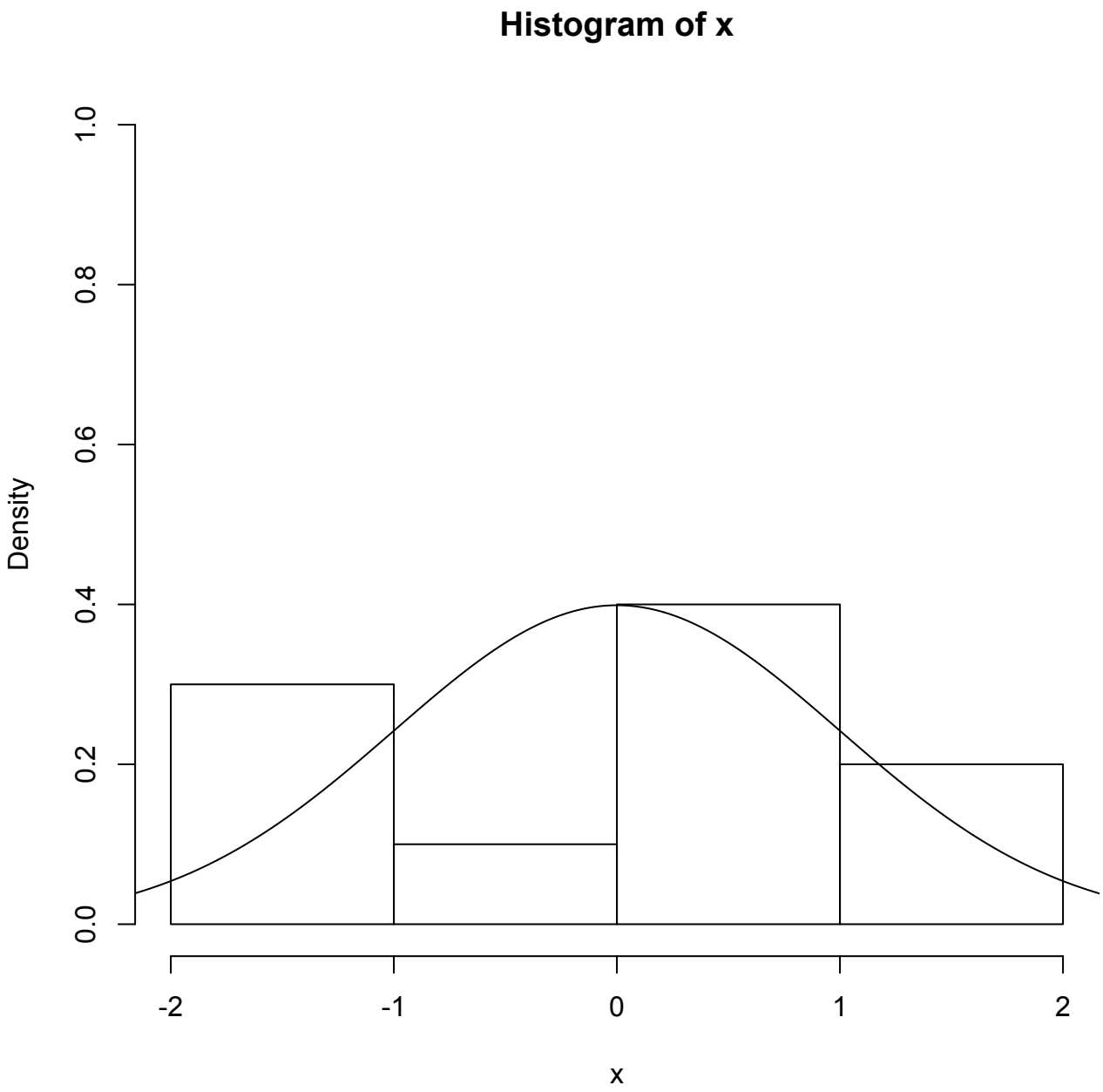
area sotto la curva fra -0.5 e +1.5 DS:

```
> pnorm(1.5) - pnorm(-0.5)
[1] 0.6246553
```

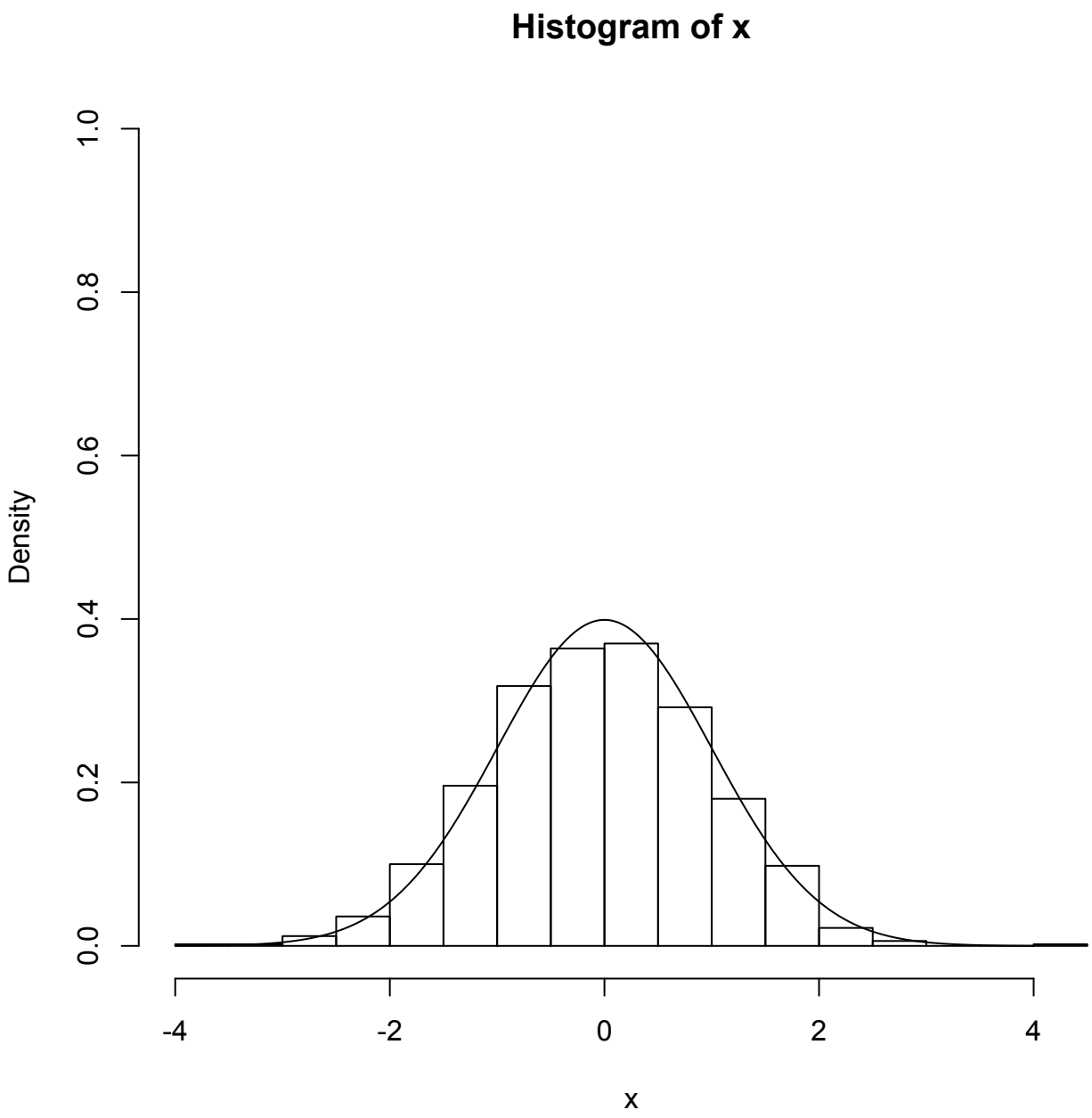


```
> qnorm(0.5)
[1] 0
> qnorm(0.5, mean = 10)
[1] 10
> qnorm(0.975)
[1] 1.959964
> qnorm(0.99)
[1] 2.326348
> qnorm(0.01)
[1] -2.326348
```

```
> x <- rnorm(10)

> hist(x, prob = TRUE, ylim = c(0, 1))

> xx <- seq(-3, 3, by = 0.01)

> lines(xx, dnorm(xx))
```


Histogram of x
Density
0.0
0.2
0.4
0.6
0.8
1.0
-2
-1
0
1
2
x

```
> x <- rnorm(1000)

> hist(x, prob = TRUE, ylim = c(0, 1))

> xx <- seq(-4, 4, by = 0.01)

> lines(xx, dnorm(xx))
```


Histogram of x
Density
0.0
0.2
0.4
0.6
0.8
1.0
-4
-2
0
2
4
x

“legge” dei grandi numeri

al crescere di n

quanto più n (la grandezza del campione) aumenta, tanto più la distribuzione del campione tende ad assomigliare alla distribuzione della popolazione

al crescere di n

M -> μ

la media di un campione è uno stimatore privo di errore sistematico della media della popolazione

il "valore atteso" per M è μ

```
> mean(rnorm(3, mean = 10))
[1] 10.19373

> mean(rnorm(3, mean = 10))
[1] 8.981192

> mean(rnorm(3, mean = 10))
[1] 9.745703

> mean(rnorm(3, mean = 10))
[1] 10.26496
```

```
> mean(rnorm(30, mean = 10))
[1] 10.23708

> mean(rnorm(30, mean = 10))
[1] 9.910758

> mean(rnorm(30, mean = 10))
[1] 9.915046

> mean(rnorm(30, mean = 10))
[1] 10.14261
```

```
> mean(rnorm(300, mean = 10))
[1] 10.01276

> mean(rnorm(300, mean = 10))
[1] 10.06409

> mean(rnorm(300, mean = 10))
[1] 10.11354

> mean(rnorm(300, mean = 10))
[1] 10.06112
```

legge dei piccoli numeri

sottostimare l'effetto dell'errore casuale su campioni piccoli

detta anche "fallacia del campione insufficiente" o "tendenza alla generalizzazione prematura"

con un campione piccolo

la forma della distribuzione può essere molto diversa dalla popolazione

M non ha errore sistematico ma può essere anche molto diversa da μ

DS ha errore sistematico: tende ad essere minore di σ

```
> mean(rnorm(5, mean = 10))
[1] 11.0353

> x <- rnorm(5)
> sd.noncorretta <- sqrt(sum(x^2)/5)

> sd.noncorretta
[1] 0.7872136

> sd(x)
[1] 0.8038325
```

la distribuzione binomiale

caratteristiche

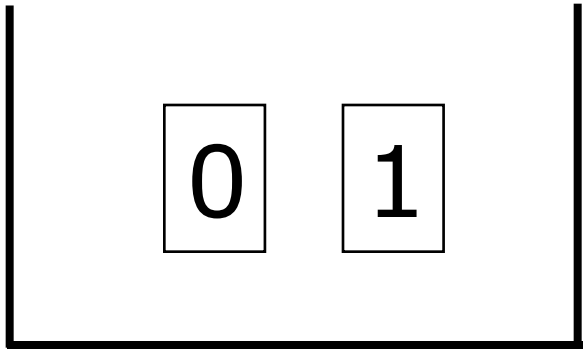
- **distribuzione discreta del *numero di successi* in una sequenza di *n* prove indipendenti**
- **le prove hanno solo due esiti mutualmente esclusivi (successo o fallimento)**
- **la probabilità di un successo è *p* e la probabilità di un fallimento è $q = 1 - p$**

esempi

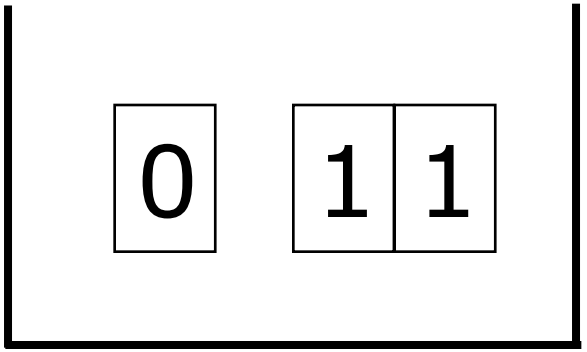
numero di volte in cui esce testa in 10 lanci di una moneta onesta ($n = 10$, $p = 1/2 = 0.5$)

numero di risposte corrette in un esame con 20 risposte a scelta multipla a 5 alternative ($n = 20$, $p = 1/5 = 0.2$), da parte di uno studente che risponda solo a caso

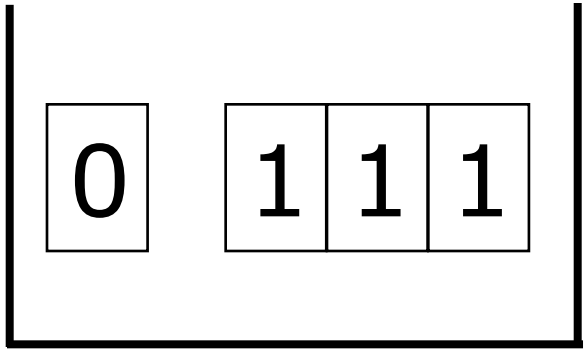
urne



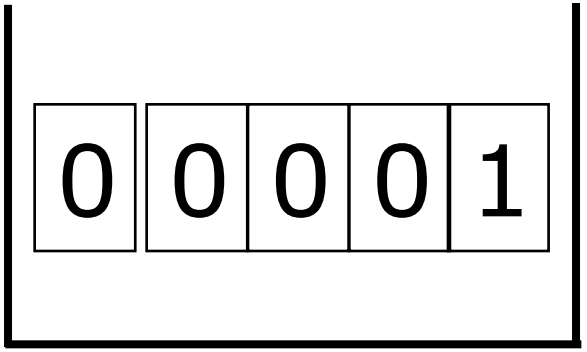
p = 1/2



p = 2/3



p = 3/4



p = 1/5

funzione di probabilità

$$f(k;n,p) = \Pr(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

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

combinazioni di n elementi in sottoinsiemi di k elementi (indipendentemente dall'ordine)

ripasso di calcolo combinatorio

$n! = n * (n - 1) * (n - 2) * \ldots 1$

<u>permutazioni</u> di n elementi: $n!$

<u>disposizioni</u> di n elementi a k a k:

$D(n, k) = n! / (n - k)!$

<u>combinazioni</u> di n elementi a k a k:

$C(n, K) = n! / [k! (n - k)!]$

funzione di probabilità

$$f(k;n,p) = \Pr(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

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

combinazioni di n elementi in sottoinsiemi di k elementi (indipendentemente dall'ordine)

n = 3, x = 1, p = 0.2

1 0 0 0.2 x 0.8 x 0.8 = 0.128

0 1 0 0.8 x 0.2 x 0.8 = 0.128

0 0 1 0.8 x 0.8 x 0.2 = 0. 128

$$pr(X = 1) = \binom{3}{1} 0.2^1 \times 0.8^2$$

= 3 x 0.128 = 0. 384

```
> dbinom(1, 3, 0.2)
[1] 0.384
```

forma

dipende da n e da k, se n è grande tende alla normale

la media è *np*

la deviazione standard è *np*(1 - *p*)

lavorare con la distribuzione binomiale

dbinom()

densità discreta

pbinom()

probabilità cumulativa da 0 a x successi

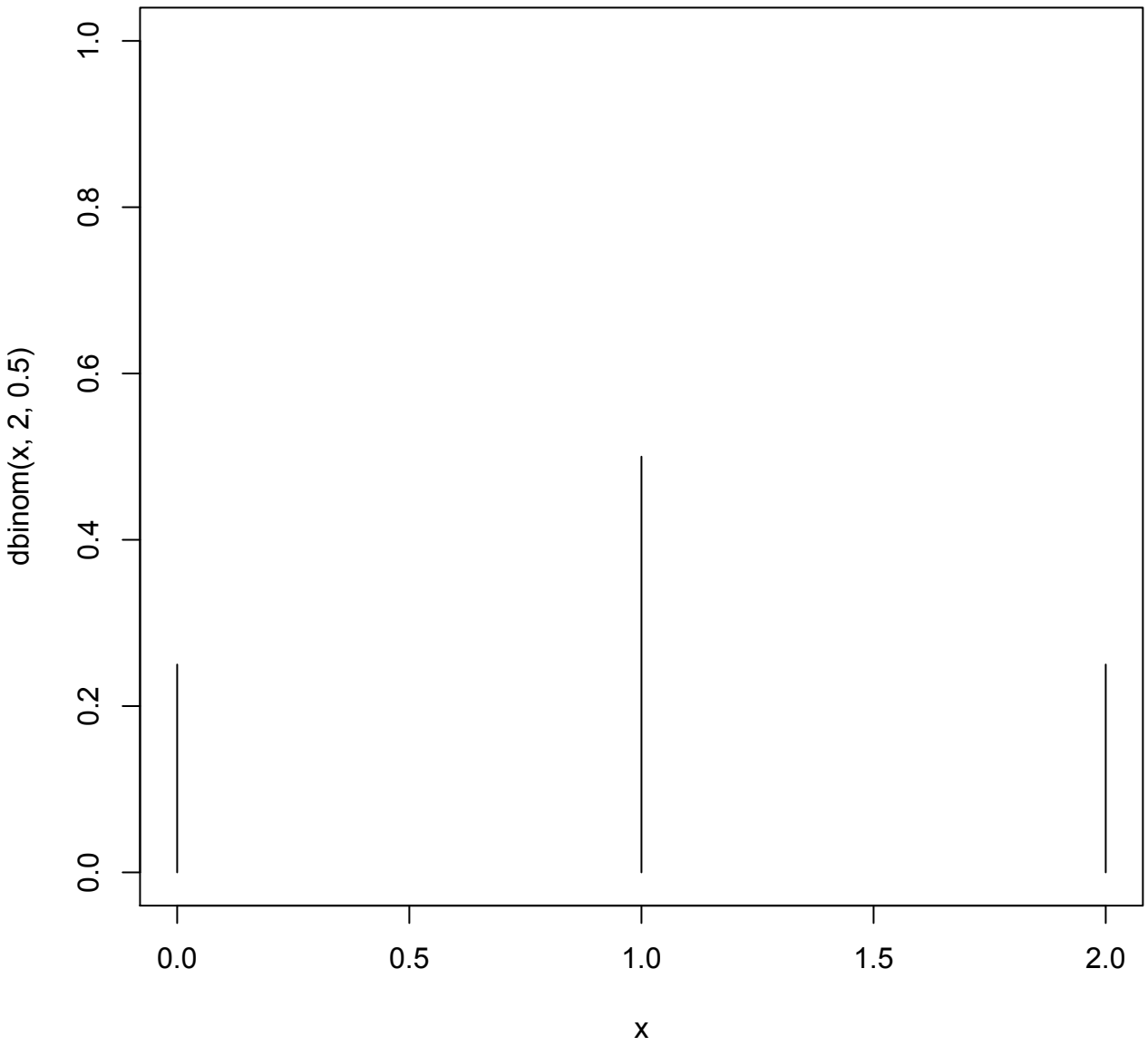
qbinom()

inversa di pbinom()

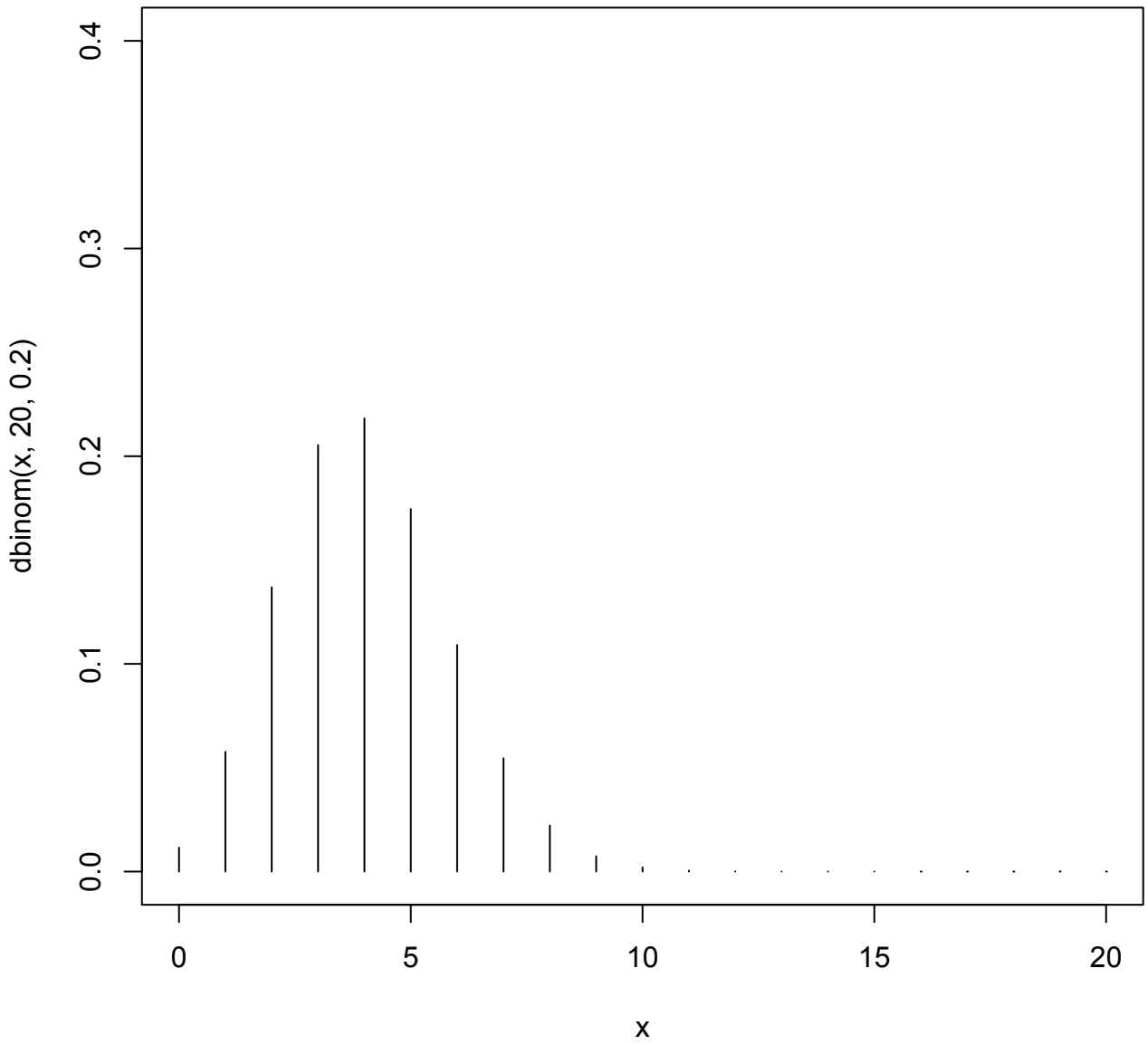
rbinom()

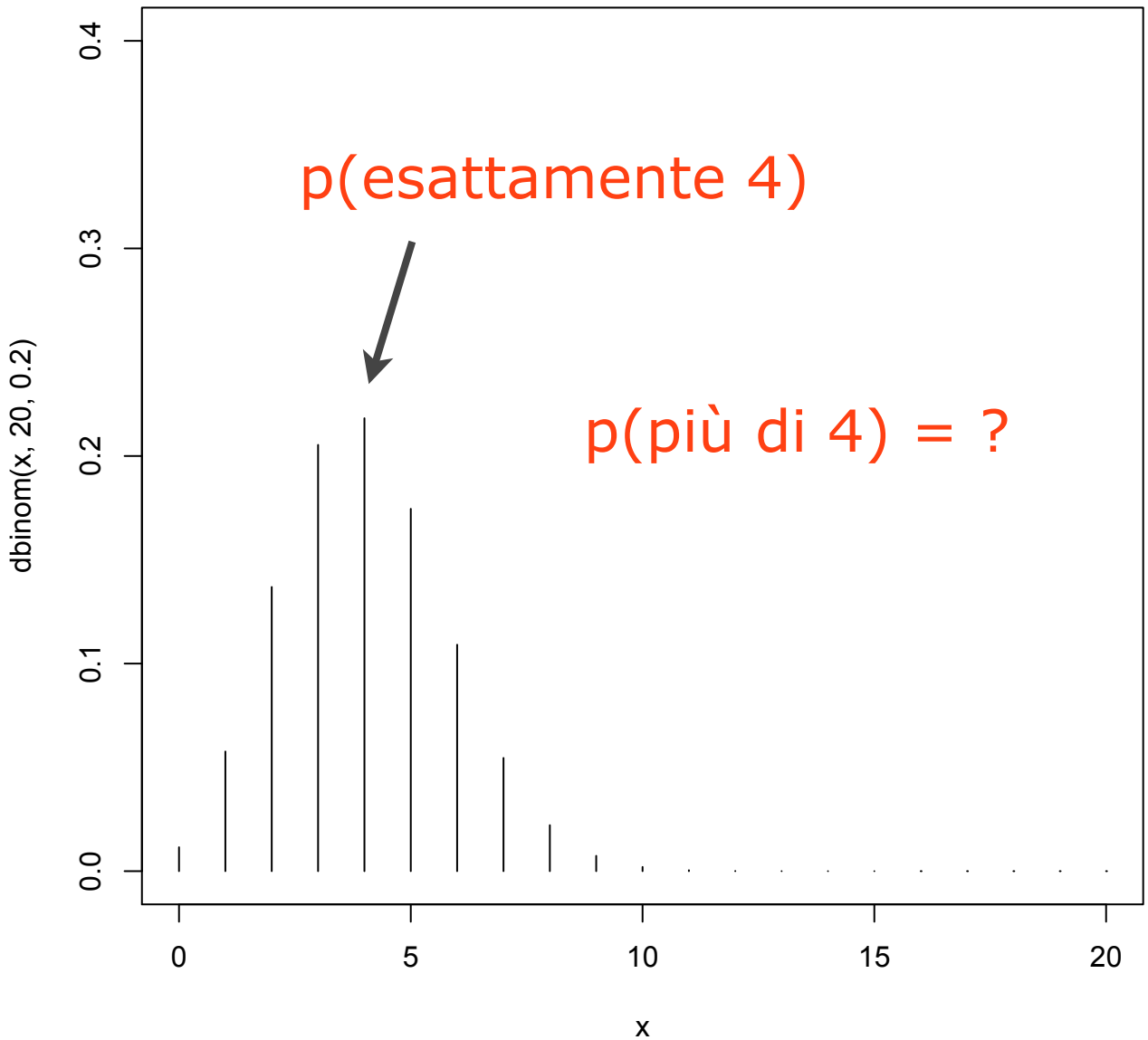
campione casuale di grandezza n

```
> dbinom(x = 0, size = 2, prob = 0.5)
[1] 0.25
> dbinom(1, 2, 0.5)
[1] 0.5
> dbinom(2, 2, 0.5)
[1] 0.25

> x <- 0:2
> plot(x, dbinom(x, 2, 0.5), type = "h",
  ylim = c(0, 1))
```


dbinom(x, 2, 0.5)
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.5
1.0
1.5
2.0
x

```
> dbinom(10, 20, 0.2)
[1] 0.002031414

> dbinom(4, 20, 0.2)
[1] 0.2181994

> x <- 0:20
> plot(x, dbinom(x, 20, 0.2), type = "h",
  ylim = c(0, 0.4))
```

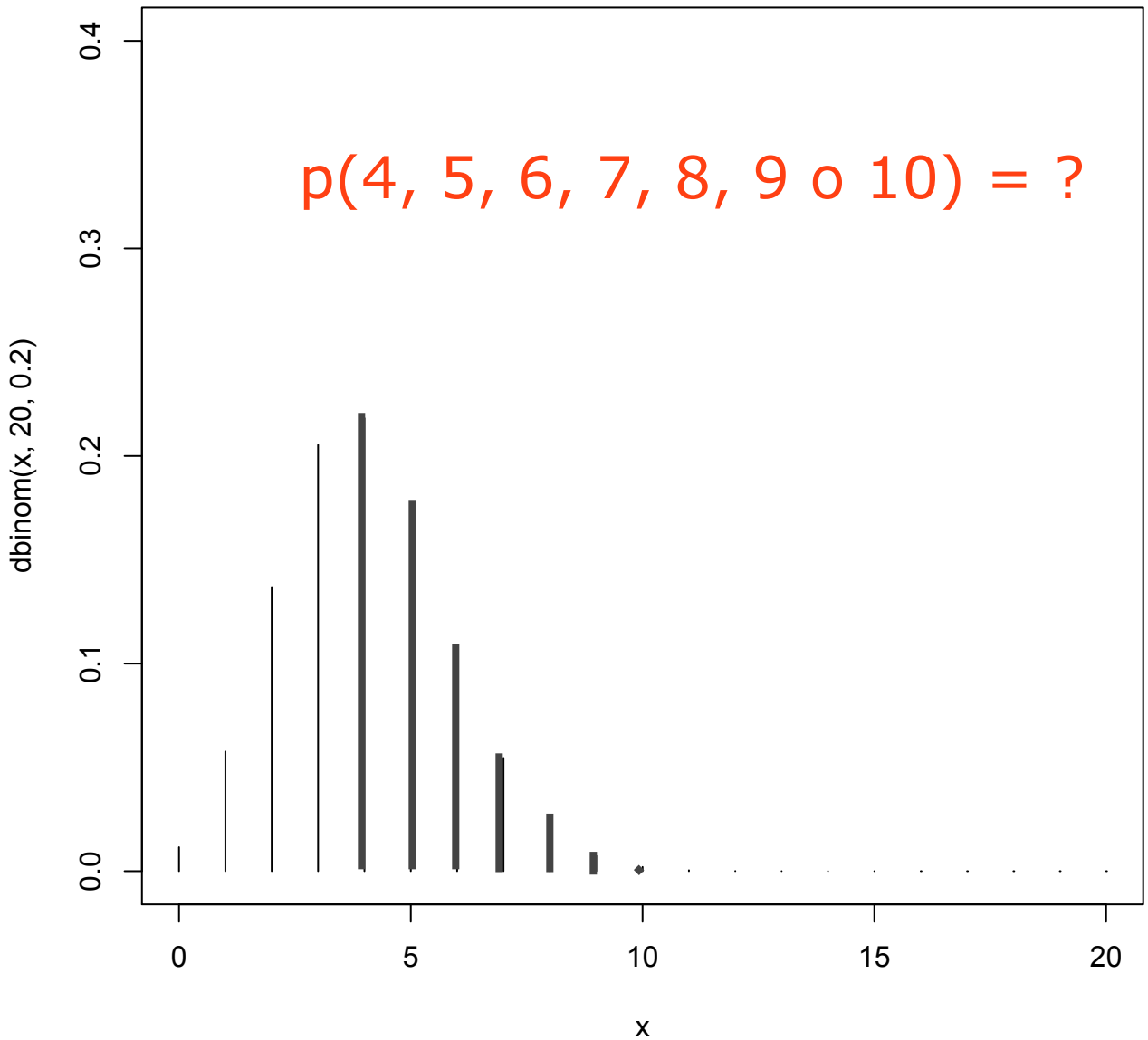

dbinom(x, 20, 0.2)
0.0
0.1
0.2
0.3
0.4
0
5
10
15
20
x


p(esattamente 4)
p(più di 4) = ?
dbinom(x, 20, 0.2)
0.0
0.1
0.2
0.3
0.4
0
5
10
15
20
x

```
> x <- 0:20
> m <- matrix(dbinom(x, 20, 0.2), nrow = 1)
> m
        [,1]       [,2]      [,3]      [,4]      [,5]
[1,] 0.01152922 0.05764608 0.1369094 0.2053641 0.2181994
        [,6]      [,7]      [,8]      [,9]      [,10]
[1,] 0.1745595 0.1090997 0.05454985 0.02216088
  0.007386959
        [,11]       [,12]        [,13]        [,14]
[1,] 0.002031414 0.0004616849 8.656592e-05 1.331783e-05
        [,15]       [,16]       [,17]        [,18]
[1,] 1.664729e-06 1.664729e-07 1.30057e-08 7.65041e-10
        [,19]       [,20]        [,21]
[1,] 3.18761e-11 8.388608e-13 1.048576e-14
```

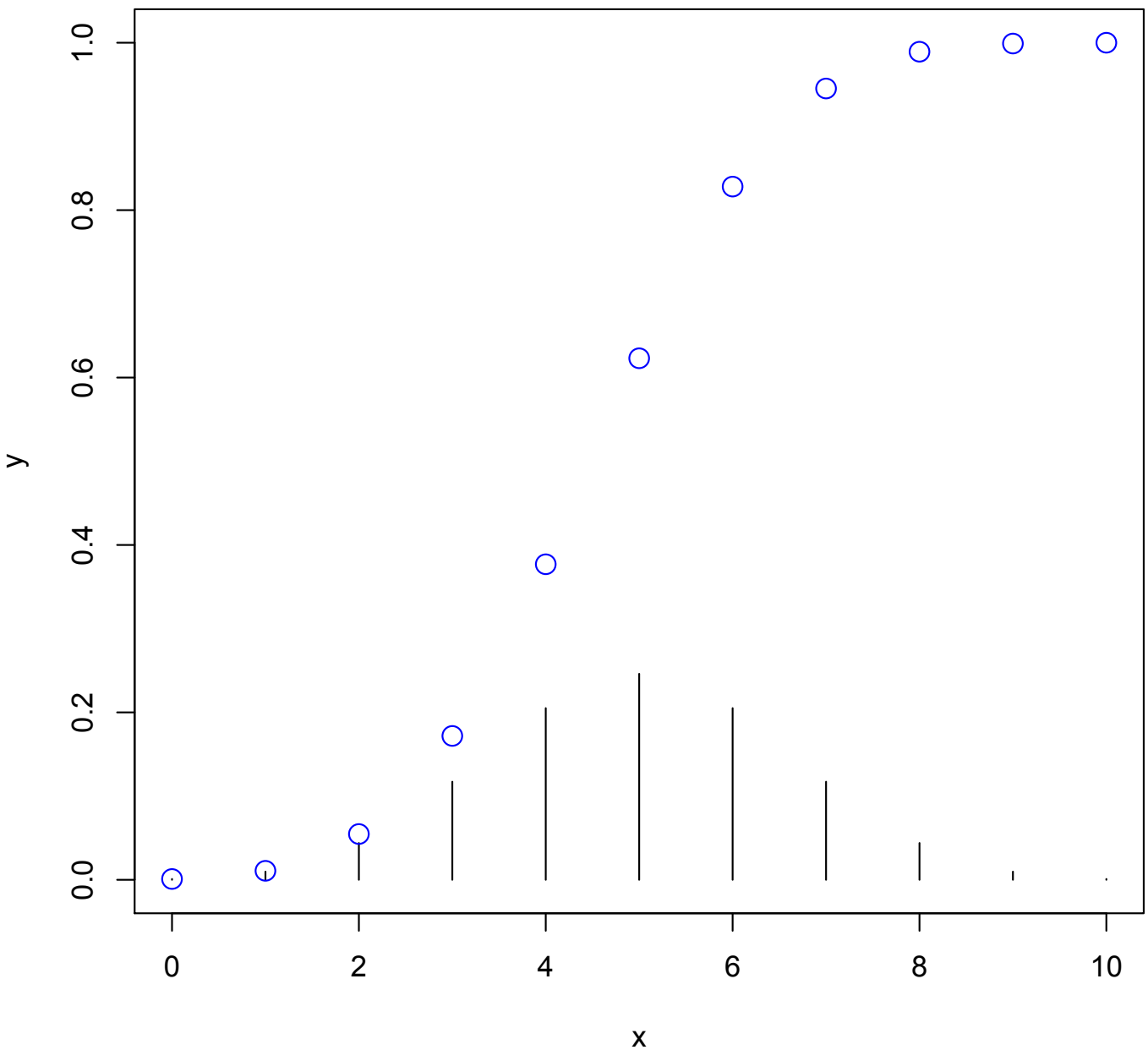
```
> colnames(m) <- x
> m
           0          1          2          3          4
[1,] 0.01152922 0.05764608 0.1369094 0.2053641 0.2181994
          5         6         7         8          9
[1,] 0.1745595 0.10909 97 0.05454985 0.02216088 0.007386959
           10          11            12                  13
[1,] 0.002031414 0.0004616849 8.656592e-05 1.331783e-05
            14          15         16                17
[1,] 1.664729e-06 1.664729e    -07 1.30057e-08 7.65041e-10
            18          19              20
[1,] 3.187671e-11 8.388608e-13 1.048576e-14
```

```
> sum(m[1:5])
[1] 0.6296483

> pbinom(4, 20, 0.2)
[1] 0.6296483

> 1 - pbinom(4, 20, 0.2)
[1] 0.3703517

> sum(m[6:20])
[1] 0.3703517
```


p(4, 5, 6, 7, 8, 9 o 10) = ?
dbinom(x, 20, 0.2)
0.0
0.1
0.2
0.3
0.4
0
5
10
15
20
x

```
> sum(dbinom(4:10, 20, 0.2))
[1] 0.587987737

> pbinom(10, 20, 0.2) - pbinom(3,
  20, 0.2)
[1] 0.5879877
```

```
> x <- 0:10
> y <- dbinom(x, 10, 0.5)

> plot(x, y, type = "h", ylim = c(0, 1))
> points(x, pbinom(x, 10, 0.5),
           col = "blue", cex = 1.5)
```

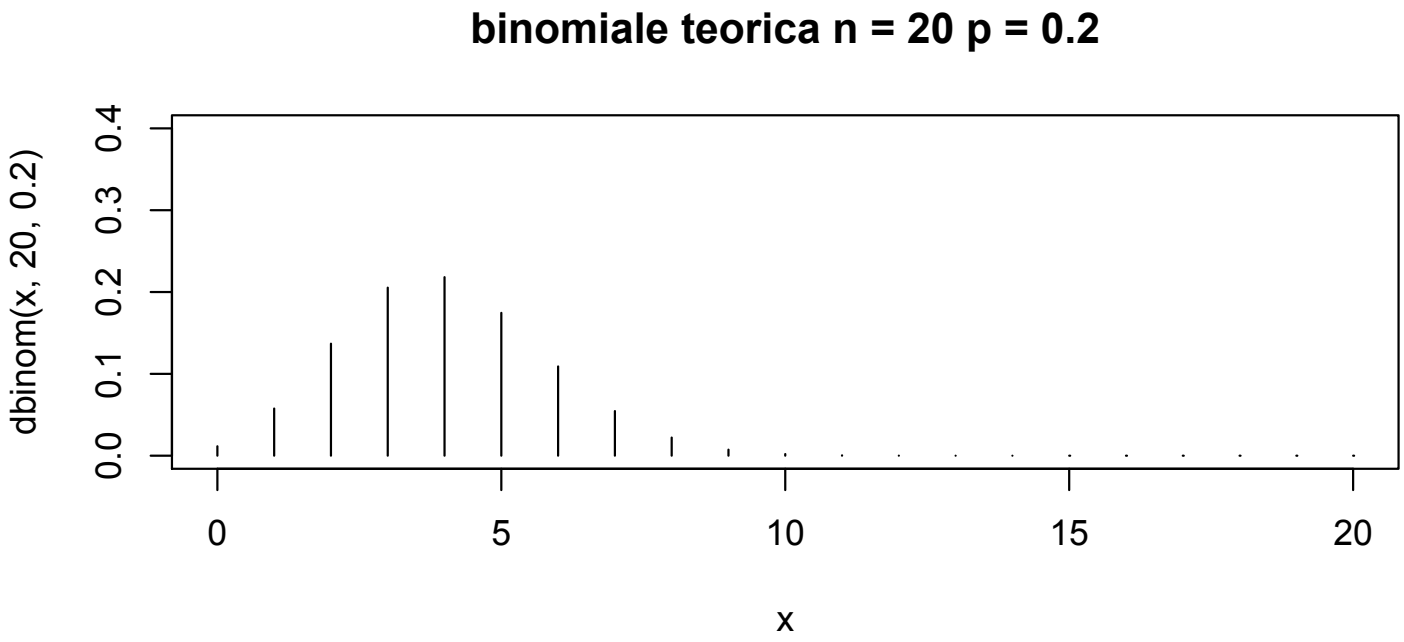

y
0.0
0.2
0.4
0.6
0.8
1.0
0
2
4
6
8
10
x

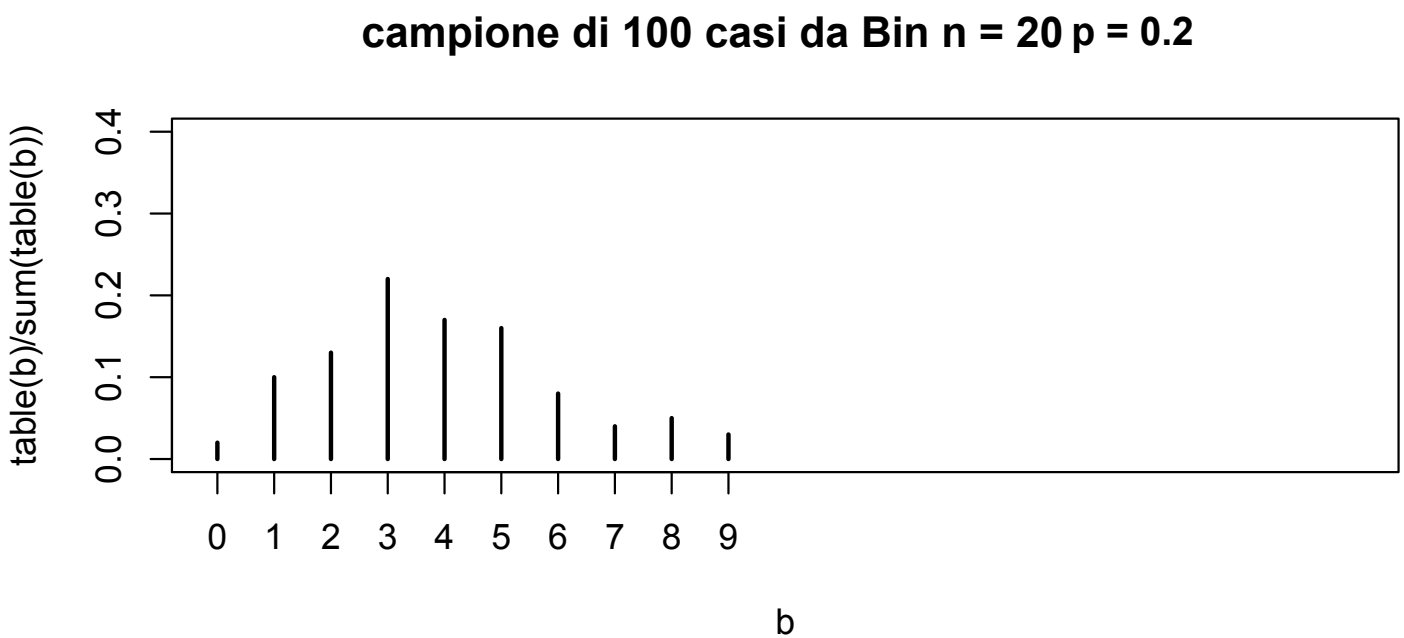
```
> pbinom(3, 20, 0.2)
[1] 0.4114489

> qbinom(0.4114489, 20, 0.2)
[1] 4

> qbinom(0.411448**8**, 20, 0.2)
[1] 3
```

```
> b <- rbinom(100, 20, 0.2)

> b
 [1] 4 2 6 5 3 4 3 7 3 3 3 1 3 7 0 6 2 4 5 7
 4 4 4 2 5 5 1
[28] 6 6 3 2 6 7 3 5 5 2 3 5 4 4 6 8 2 1 2 8
 4 7 6 2 5 2 4
[55] 3 7 3 4 5 3 2 8 7 2 7 7 6 1 8 6 4 4 6 4
 1 3 6 3 3 3 4
[82] 4 3 2 7 6 3 4 5 7 1 4 5 4 6 4 3 6 6 3
```

```
> table(b)

b
 0  1  2  3  4  5  6  7  8
 1  6 12 20 20 11 15 11  4

> table(b)/sum(table(b))
b
   0    1    2    3    4    5    6    7    8
0.01 0.06 0.12 0.20 0.20 0.11 0.15 0.11 0.04
```

```
> x <- 0:20
> plot(x, dbinom(x, 20, 0.2),
        type = "h", ylim =c(0, 0.4),
        main = "binomiale teorica n = 20
        p = 0.2")

> plot(table(b)/sum(table(b)),
      type = "h", xlim = c(0, 20),
      ylim = c(0, 0.4), main = "campione
        di 100 casi da Bin n = 20 p =
        0.2")
```


binomiale teorica n = 20 p = 0.2
dbinom(x, 20, 0.2)
0.0
0.1
0.2
0.3
0.4
0
5
10
15
20
x


campione di 100 casi da Bin n = 20 p = 0.2
table(b)/sum(table(b))
0.0
0.1
0.2
0.3
0.4
0
1
2
3
4
5
6
7
8
9
b