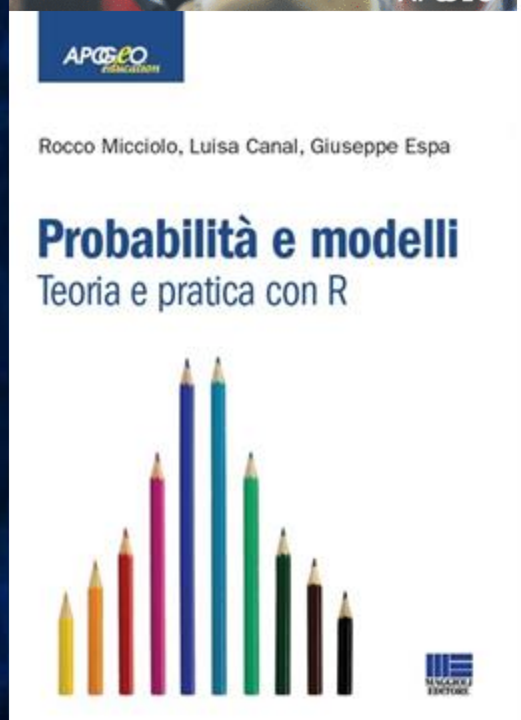
Il *package* `rmf` nell'insegnamento dei corsi di base di Statistica

Giuseppe Espa, Rocco Micciolo


APOGEO education
Problemi
ed esperimenti
di statistica con R
Giuseppe Espa
Rocco Micciolo
APOGEO


APOGEO education
Analisi esplorativa
dei dati con R
Giuseppe Espa
Rocco Micciolo
APOGEO


APOGEO education
Rocco Micciolo, Luisa Canal, Giuseppe Espa
Probabilità e modelli
Teoria e pratica con R


Rocco Micciolo
Giuseppe Espa
Luisa Canal
Ricerca con R
Metodi di inferenza statistica
APOGEO


Alan Agresti
Christine Franklin
All'interno Laboratori con R
STATISTICA
L'arte e la scienza d'imparare dai dati
A cura
e con materiali aggiuntivi di
Giuseppe Espa, Rocco Micciolo
Diego Giuliani, Maria Michela Dickson
Pearson Learning Solution
Codice di accesso a MyLab
Aula virtuale
Risorse multimediali
Test ed esercizi
Autovalutazione
Pearson eText
ALWAYS LEARNING
PEARSON


Nuovi contenuti con R
Statistica
L'arte e la scienza d'imparare dai dati
Quinta edizione
Alan Agresti, Christine Franklin, Bernhard Klingenberg
a cura di Giuseppe Espa, Rocco Micciolo,
Diego Giuliani, Maria Michela Dickson
MyLab Codice studente
Pearson

rmf

- È una libreria di funzioni sviluppata a partire dal 2004 (*Econometria ed applicazioni ai servizi sanitari*)
- È scritta esclusivamente usando funzioni preesistenti di R e si occupa in larga misura di produrre un *output*
- Permette inoltre (al docente e allo studente) di eseguire simulazioni per “dimostrare” empiricamente le regole e i metodi dell’inferenza statistica
- Consente di eseguire test di significatività e calcolare intervalli di confidenza a partire da dati sintetici (medie, d.s., proporzioni, ecc.)
- Il file .tar.gz si può scaricare andando alla pagina seguente: https://hostingwin.unitn.it/micciolo/
- Come tutti i pacchetti (*packages*) va installata una volta per tutte e richiamata all’apertura di ogni sessione di R con il comando **`library(rmf)`**

Di seguito vengono illustrate alcune funzioni implementate nel *package* **`rmf`** per calcolare statistiche descrittive, eseguire test di significatività e costruire intervalli di confidenza oggetto dei corsi di base di Statistica per le lauree triennali e per quelle a ciclo unico (Medicina e chirurgia).

Le funzioni proposte sono tutte basate su funzioni già presenti in **`R`** e si limitano, generalmente, a fornire un *output* più dettagliato.

```
> frequenze(df$residenza)
        x      n               f

+--------+------+--------------+
| Altro  |    3 |     1.132075 |
| CSI    |   14 |     5.283019 |
| Nord   |  109 |    41.132075 |
| TN     |  139 |    52.452830 |
+--------+------+--------------+
          265    100.000000
Osservazioni mancanti: 3
```

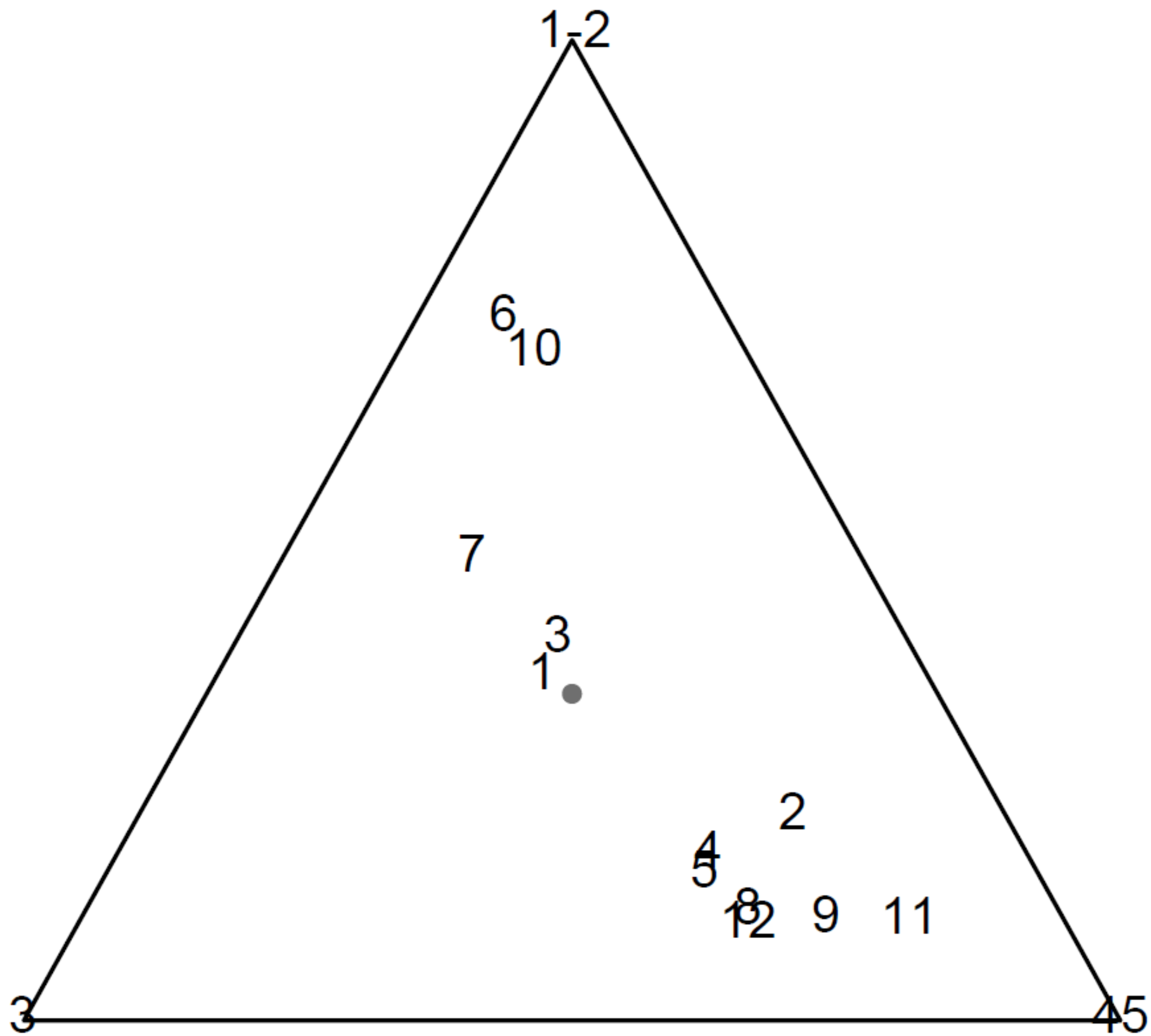
```
> frequenze(df$libro, cumul=TRUE)
  x      n            f      N              F
+---+------+------------+-----+-------------+
| 1 |    8 |   2.985075 |   8 |    2.985075 |
| 2 |   21 |   7.835821 |  29 |   10.820896 |
| 3 |   78 |  29.104478 | 107 |   39.925373 |
| 4 |   98 |  36.567164 | 205 |   76.492537 |
| 5 |   63 |  23.507463 | 268 |  100.000000 |
+---+------+------------+-----+-------------+
Osservazioni mancanti: 0
```

```
> frequenze(a,sort=TRUE,cumul=TRUE)
                           x    n            f    N           F

+----------------------+------+-------------+------+-----------+
| Base curvata         | 1987 |   31.419987 | 1987 |  31.41999 |
| Danneggiamento       | 1039 |   16.429475 | 3026 |  47.84946 |
| Piccoli segni        |  834 |   13.187856 | 3860 |  61.03732 |
| Graffi               |  442 |    6.989247 | 4302 |  68.02657 |
| Striature            |  413 |    6.530677 | 4715 |  74.55724 |
| Puntini neri         |  413 |    6.530677 | 5128 |  81.08792 |
| Segni di colpi       |  371 |    5.866540 | 5499 |  86.95446 |
| Segni di spray       |  292 |    4.617331 | 5791 |  91.57179 |
| Base ammaccata       |  275 |    4.348514 | 6066 |  95.92030 |
| Getto di inchiostro  |  258 |    4.079696 | 6324 | 100.00000 |
+----------------------+------+-------------+------+-----------+
Osservazioni mancanti: 0
```

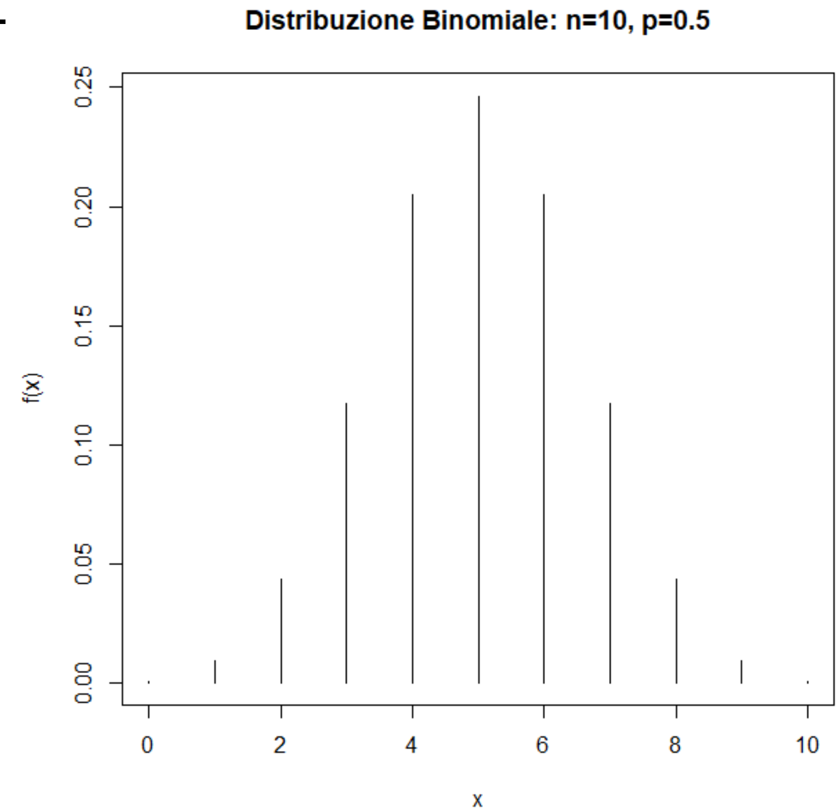
```
> tc(df$sex,df$estero)
       FR.A.  I                    TOTALE
       %RIGA  I                   DI RIGA
       %COL.  I    NO I    SI I
       ------+-------+-------+
              I    33 I    71 I       104
           F  I  31.7 I  68.3 I      67.5
              I  75.0 I  64.5 I
            -+-------+-------+
              I    11 I    39 I        50
           M  I  22.0 I  78.0 I      32.5
              I  25.0 I  35.5 I
            -+-------+-------+
    TOTALE DI      44     110         154
      COLONNA    28.6    71.4

    CHI QUADRATO =    1.126 (GRADI DI LIBERTA' =  1)
    p-value del test chi-quadrato = 0.289; p-value 'esatto' = 0.255
    CELLE CON F.A. < 5 =   0 - CELLE CON F.A. < 1 =   0

    Odds Ratio   1.00    0.61
    Reciproco    1.00    1.65
```

```
> reglin(df$altezza,df$peso)
Variabile dipendente  : $ df peso
Variabile indipendente: $ df altezza
Intercetta: -87.60099
Pendenza  : 0.878095
% variabilità spiegata: 41.80071
Numero di coppie di osservazioni: 264
Test t per l'assenza di regressione lineare: 13.71777
p-value = 1.212655e-32
```

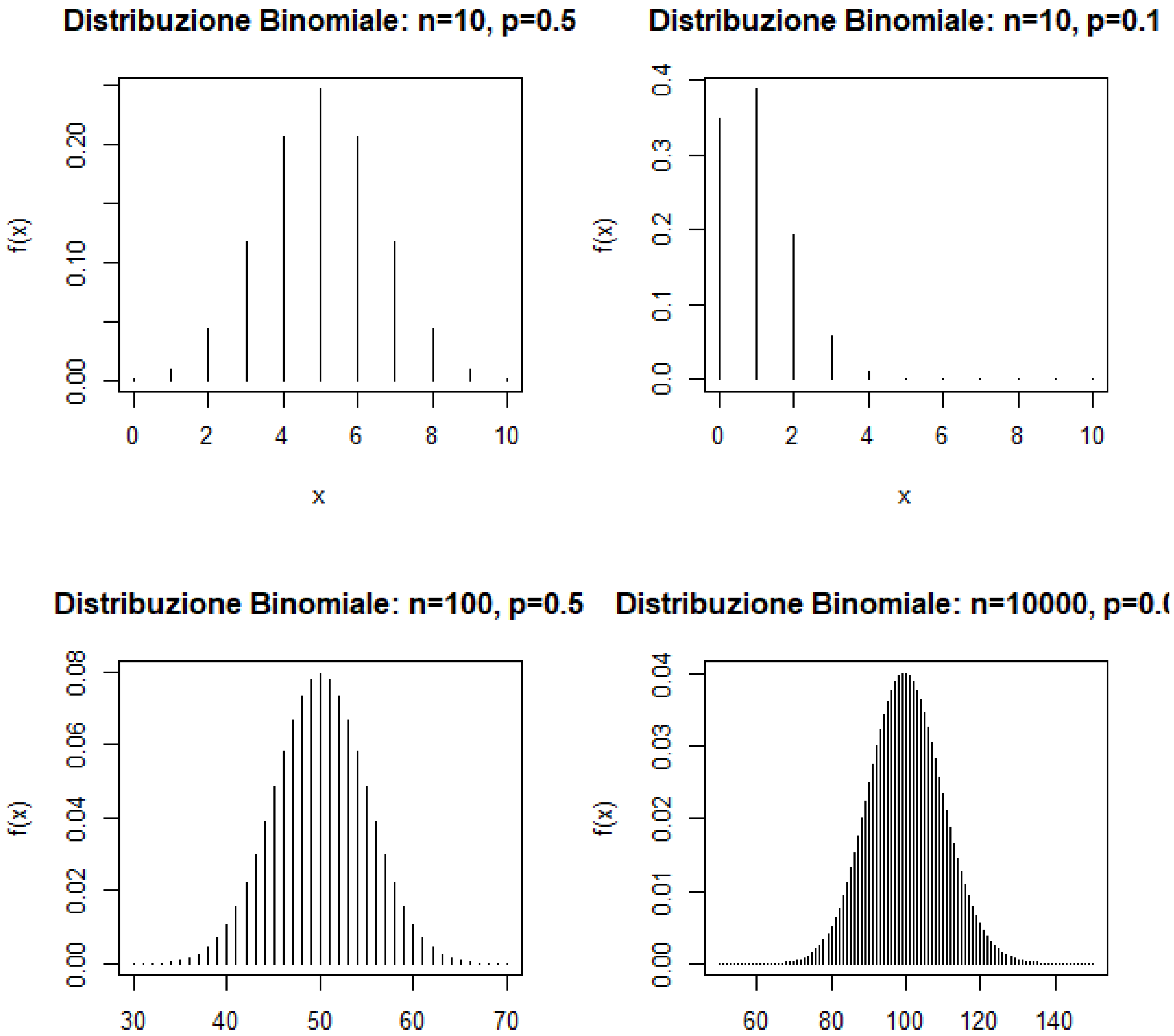
```
> triplot(ris,cex.vert=1,cex.text=1)
> centro()
```



```
> library(rmf)
> Binomiale(n=10,p=0.5)
------------------------------------
Distribuzione Binomiale
------------------------------------
Numero delle prove        : 10
Probabilita' di successo: 0.5
Valore atteso (media)     : 5
Varianza                  : 2.5
Somma delle probabilita': 1
       x           f(x)         F(x)
 [1,]  0 0.0009765625 0.0009765625
 [2,]  1 0.0097656250 0.0107421875
 [3,]  2 0.0439453125 0.0546875000
 [4,]  3 0.1171875000 0.1718750000
 [5,]  4 0.2050781250 0.3769531250
 [6,]  5 0.2460937500 0.6230468750
 [7,]  6 0.2050781250 0.8281250000
 [8,]  7 0.1171875000 0.9453125000
 [9,]  8 0.0439453125 0.9892578125
[10,]  9 0.0097656250 0.9990234375
[11,] 10 0.0009765625 1.0000000000
```

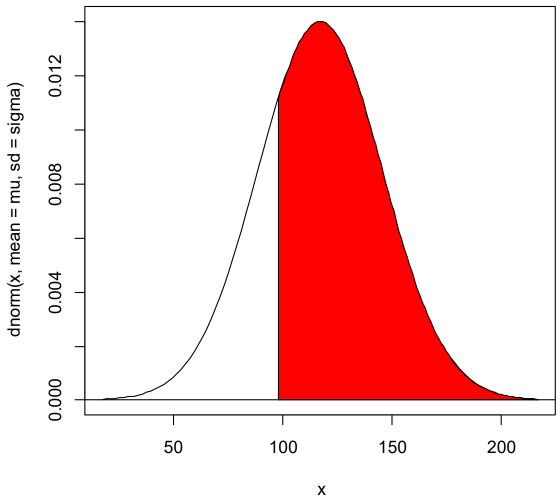


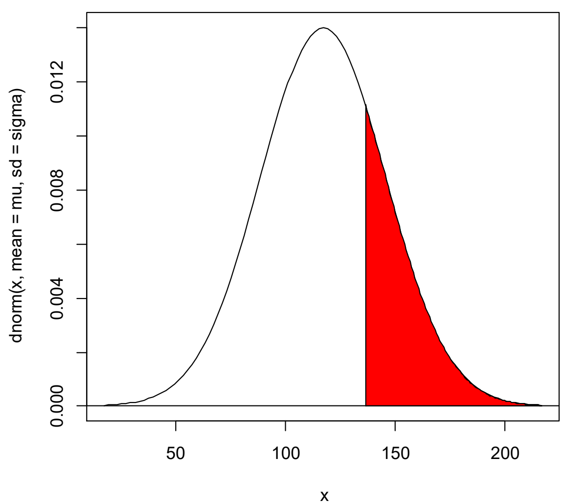
```
Binomiale(n=10,p=0.5)
Binomiale(n=10,p=0.1)
Binomiale(n=100,p=0.5,da=30,a=70)
Binomiale(n=10000,p=0.01,da=50,a=150)
```



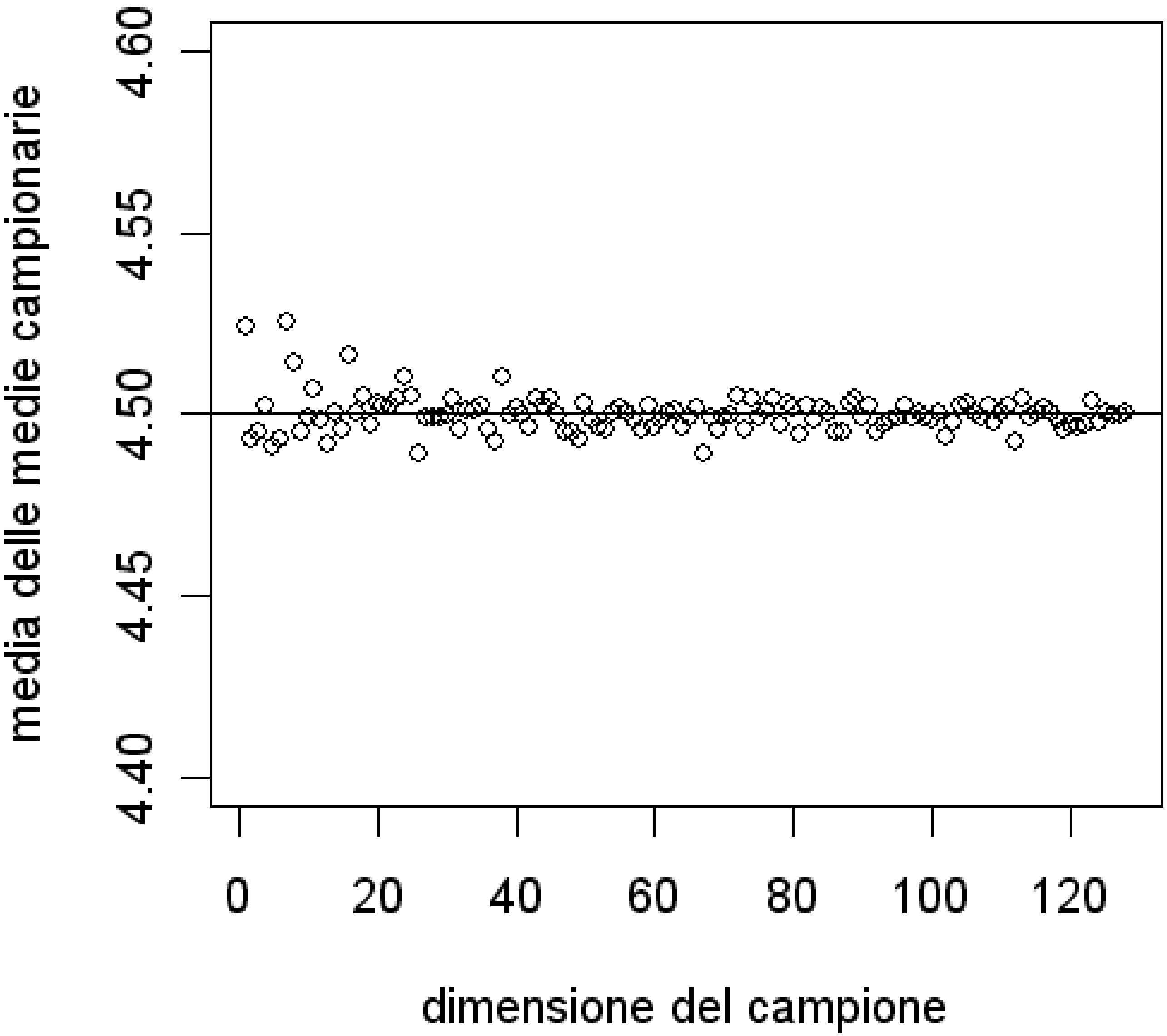
I punteggi ottenuti in un test di atteggiamento seguono una distribuzione normale con media 117 e deviazione standard 28.5. (a) Quale è la probabilità di ottenere un punteggio superiore a 131? (b) Quale punteggio è superato dal 75% dei soggetti sottoposti al test? (c) Quale punteggio è superato dal 25% dei soggetti sottoposti al test? (d) Se si considera un campione di 100 soggetti sottoposti al test, quale è la probabilità che il punteggio medio sia superiore a 128?

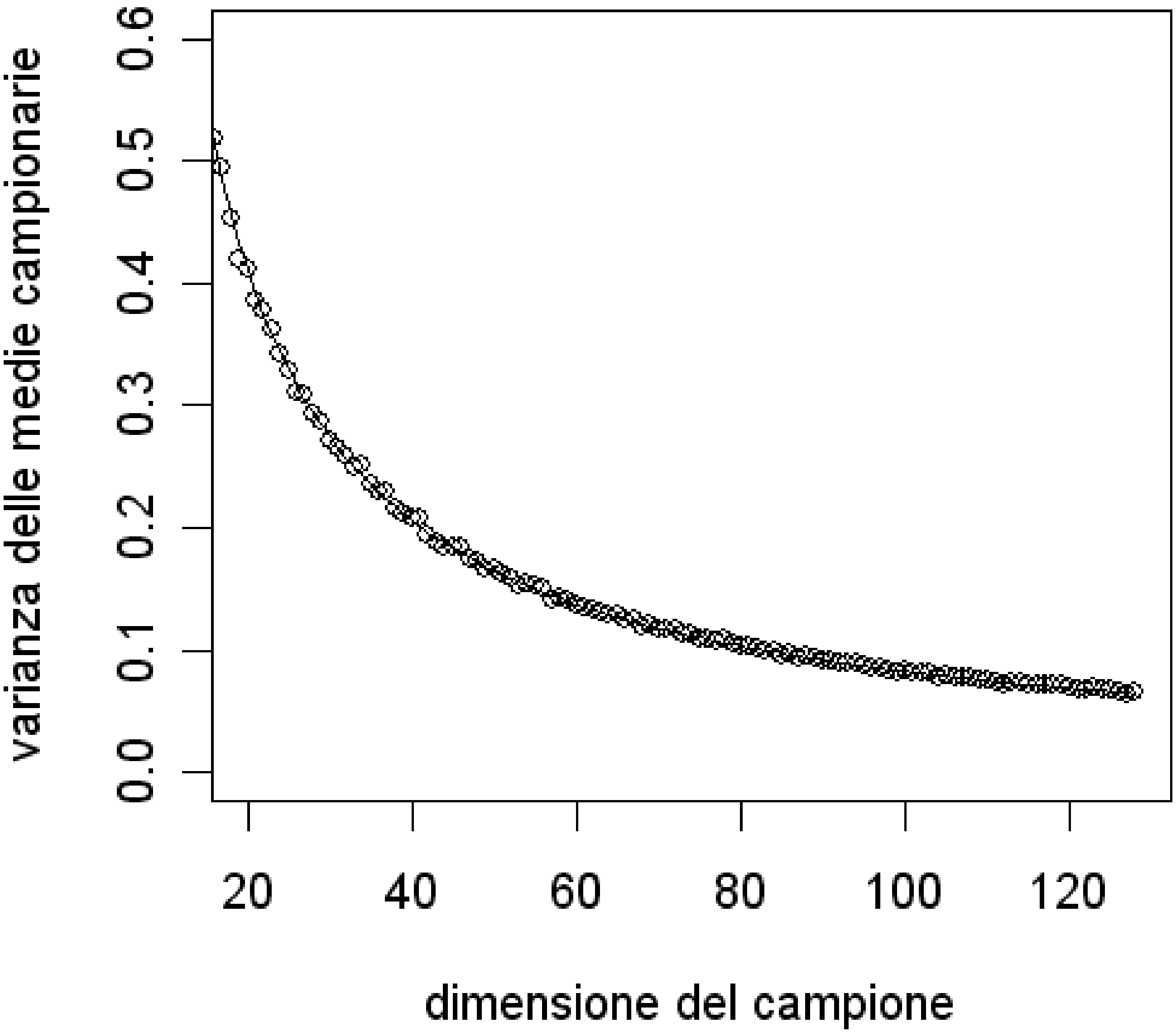
```
> 1 - pnorm(131, 117, 28.5)
[1] 0.3116326

> qnorm(0.25)
[1] -0.6744898
> qnorm(0.25, 117, 28.5)
[1] 97.77704

> qnorm(0.75)
[1] 0.6744898
> qnorm(0.75, 117, 28.5)
[1] 136.223
```

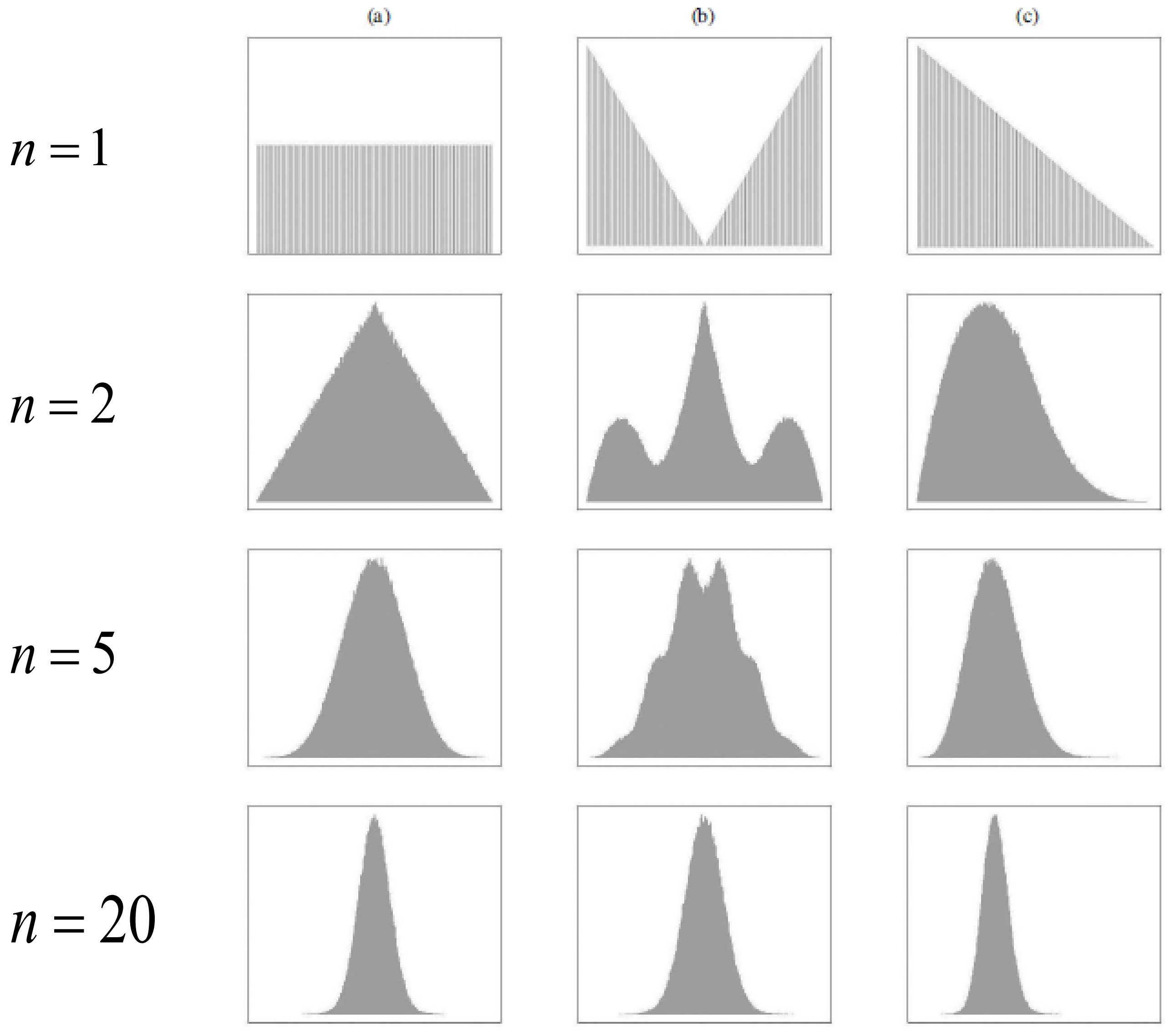




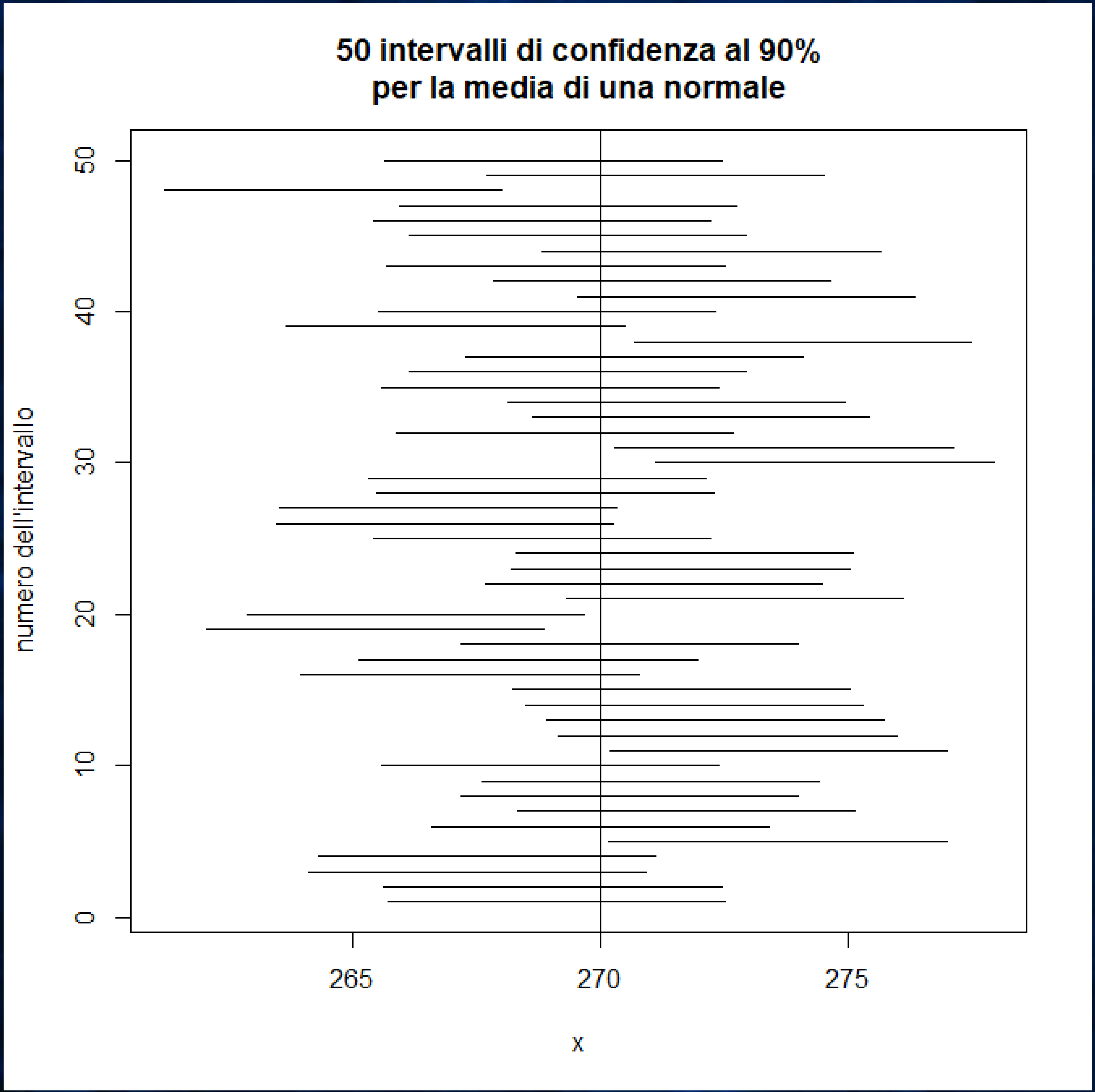
```
library(rmf)

pop <- c(0:9)

xm <- sim(n=1,pop,nrep=10000)
xm <- sim(n=2,pop,nrep=10000)
xm <- sim(n=4,pop,nrep=10000)
xm <- sim(n=8,pop,nrep=10000)
xm <- sim(n=16,pop,nrep=10000)
xm <- sim(n=32,pop,nrep=10000)
xm <- sim(n=64,pop,nrep=10000)
xm <- sim(n=128,pop,nrep=10000)

Simulazione di 10000 campioni, ciascuno di dimensione 128,
estratti da una distribuzione empirica
con media 4.5 e varianza 8.25

La media     delle medie e'   4.500105
La varianza delle medie e'   0.06328038
La d.s.      delle medie e'   0.2515559
```


media delle medie campionarie
4.40
4.45
4.50
4.55
4.60
0
20
40
60
80
100
120
dimensione del campione

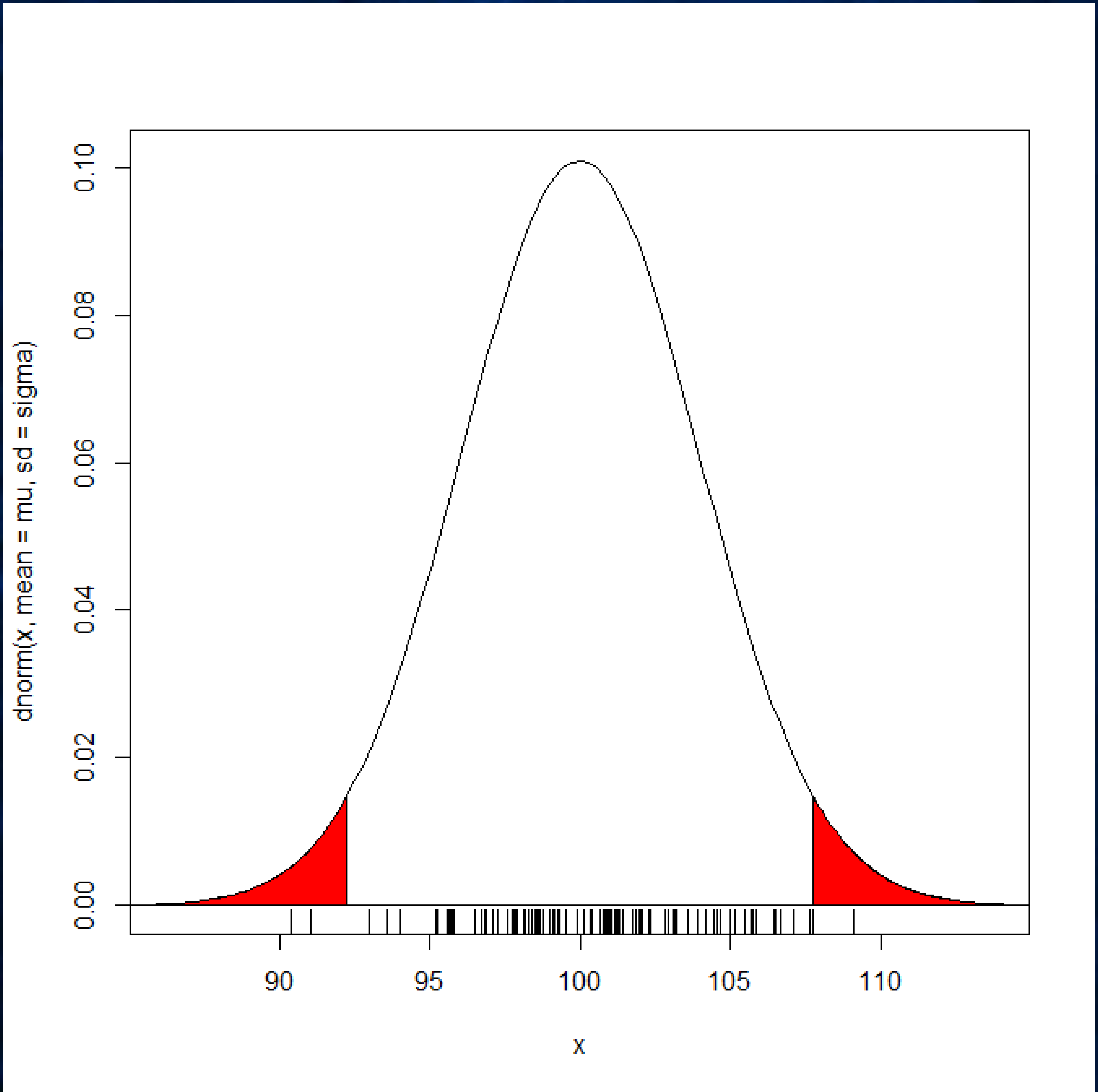

varianza delle medie campionarie
0.0
0.1
0.2
0.3
0.4
0.5
0.6
20
40
60
80
100
120
dimensione del campione


(a)
(b)
(c)
$n = 1$
$n = 2$
$n = 5$
$n = 20$

```
> library(rmf)
>
> set.seed(654321)
> ic <- simIC(n=840, mu=270, sigma=60, conf=0.9, nrep=50)
> summary(ic)
Simulazione di 50 intervalli di confidenza al 90%
su campioni di dimensione 840 estratti da una normale
con media 270 e deviazione standard 60
Numero di intervalli che cadono a sinistra di  270  : 3
Numero di intervalli che cadono a destra   di  270  : 5
Numero di intervalli che contengono il valore  270  : 42
Percentuale di intervalli che contengono       270  : 84
> plot(ic)
```

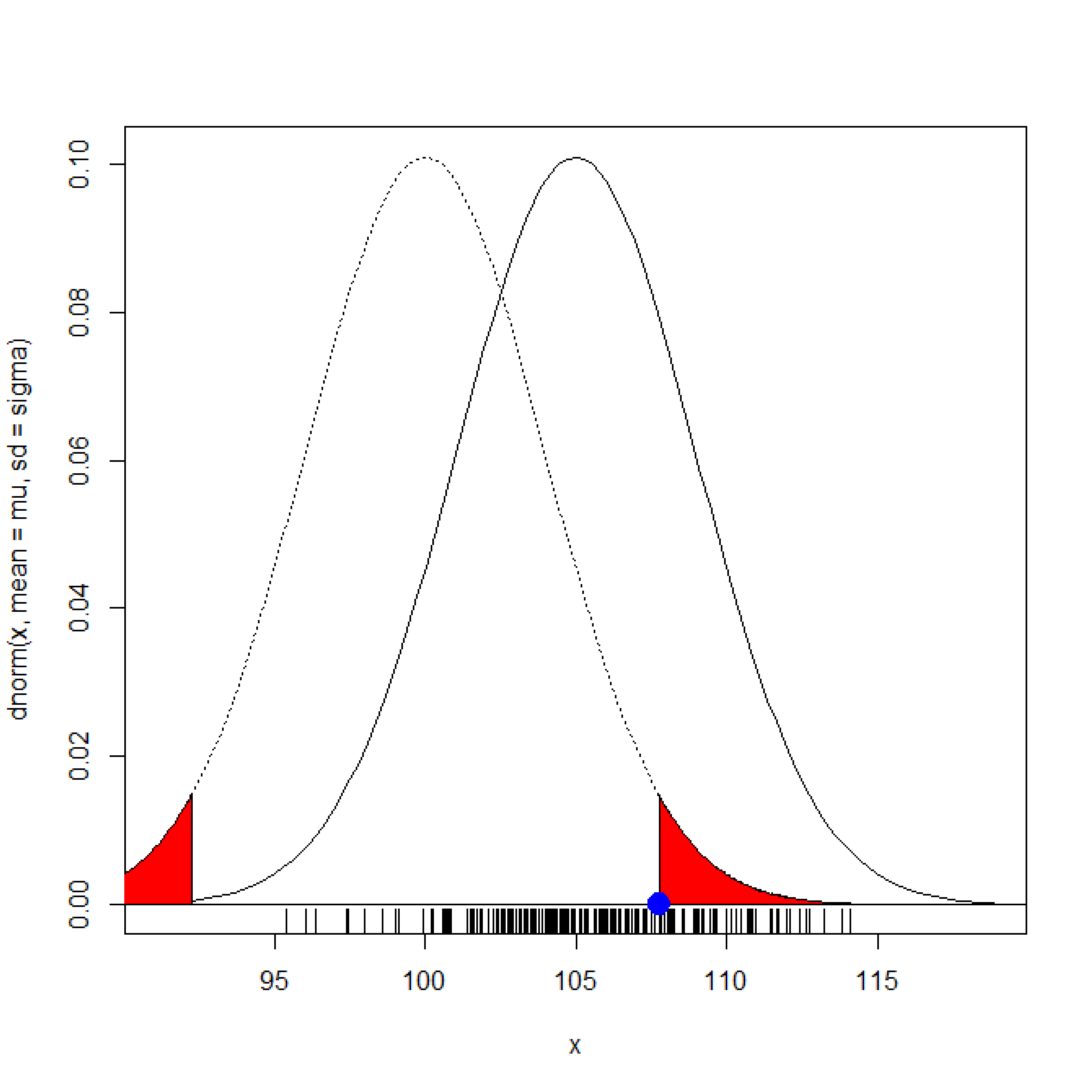

50 intervalli di confidenza al 90%
per la media di una normale
numero dell'intervallo
0
10
20
30
40
50
265
270
275
x

```
> library(rmf)
>
> mpop <- 100; dspop <- 25; dimc <- 40
> set.seed(123456)
> tmp <- simTest(n=dimc,mu0=mpop,sigma=dspop,alpha=0.05,nrep=100)
> summary(tmp)
Simulazione di 100 campioni, ciascuno di dimensione 40
per verificare l'ipotesi nulla mu = 100
rispetto ad una alternativa bilaterale
ad un livello di significativita' alpha = 0.05
quando l'ipotesi nulla e' vera.
La deviazione standard della popolazione e' 25
e l'errore standard e' 3.952847
Regione di rifiuto: medie inferiori a 92.25  o superiori a 107.75
Numero di test non significativi :              96
Numero di test significativi (coda sinistra): 2
Numero di test significativi (coda destra):   2
Percentuale di test significativi:            4
> plot(tmp)
```



dnorm(x, mean = mu, sd = sigma)
0.00
0.02
0.04
0.06
0.08
0.10
90
95
100
105
110
x

```
> library(rmf)
>
> mpop <- 105; dspop <- 25; dimc <- 40
> set.seed(123456)
> tmp <- simTest(n=dimc,mu0=100,mu1=mpop,
sigma=dspop,alpha=0.05,nrep=200)
> summary(tmp)
Simulazione di 200 campioni, ciascuno di dimensione 40
per verificare l'ipotesi nulla mu = 100
rispetto ad una alternativa bilaterale
ad un livello di significativita' alpha = 0.05
quando l'ipotesi nulla e' falsa e la media vera e' 105
La deviazione standard della popolazione e' 25
e l'errore standard e' 3.952847
Regione di rifiuto: medie inferiori a 92.25  o superiori a 107.75
Numero di test non significativi :             148
Numero di test significativi (coda sinistra): 0
Numero di test significativi (coda destra):   52
Percentuale di test significativi:            26
> plot(tmp)
```



dnorm(x, mean = mu, sd = sigma)
0.00
0.02
0.04
0.06
0.08
0.10
95
100
105
110
115
x

Esempio 1.1 La *Federal Trade Commission* (*FTC*) sottopone periodicamente a verifica le dichiarazioni dei produttori a proposito dei prodotti che commercializzano. In particolare la *FTC* vuole verificare se i barattoli di caffè venduti dalla *Hilltop Coffee* contengono effettivamente 3 libbre di caffè. Al riguardo la *FTC* seleziona un campione di $n = 36$ barattoli di caffè *Hilltop* per il quale la media campionaria $\overline{y} = 2.92$ libbre. Sapendo da studi precedenti che la deviazione standard σ della popolazione può essere considerata nota e pari a 0.18 libbre, il risultato campionario è compatibile con quanto dichiarato nelle etichette dalla *Hilltop*? (Anderson, Sweeney, Williams, 2011).

```
> test.z(media=2.92,sigma=0.18,n=36,mu0=3,conf=0.95)
media = 2.92  d.s. = 0.18  es = 0.03  n = 36
Intervallo di confidenza al  95% per la media:
da  2.861201  a  2.978799

Ipotesi nulla: mu = 3 media = 2.92  ES = 0.03  n = 36
Regione di non rifiuto per l'ipotesi nulla (alpha=0.05):
da  2.941201  a  3.058799
Test z: -2.666667  p-value = 0.007660761
```

In un esperimento un gruppo di 9 uomini ha bevuto mezza bottiglia di vino rosso ciascuno per due settimane. Il livello di polifenoli nel loro sangue è stato misurato prima e dopo le due settimane.
Le differenze percentuali sono risultate le seguenti: 3.5, 8.1, 7.4,4.0,0.7,4.9,8.4,7.0,5.5
Calcolare un intervallo di confidenza al 90% per il cambiamento percentuale medio nel livello di polifenoli nel sangue.

```
> vino <-  c(3.5,8.1,7.4,4.0,0.7,4.9,8.4,7.0,5.5)

> test.t1c(dati=vino,conf=0.9)

media = 5.5  d.s. = 2.516943  es = 0.8389809
n = 9  t critico = 1.859548
Intervallo di confidenza al 90% per la media:
da 3.939875 a 7.060125
```

```
> library(rmf)

> test.t1c(5.5, 2.516943, 9, mu0 = 7.5)

media = 5.5  d.s. = 2.516943  ES = 0.838981
n = 9  t critico = 2.306004
Intervallo di confidenza al 95% per la media:
da 3.565306 a 7.434694

Ipotesi nulla: mu = 7.5
Test t: -2.383844  g.l. = 8
p-value = 0.04427921
```

Esiste una differenza fra gli studenti universitari che svolgono attività di volontariato e quelli che si dedicano solo allo studio? Una ricerca ha analizzato i dati di 57 studenti che hanno partecipato ad alcune attività di volontariato e di altri 17 studenti che invece non hanno mai svolto servizi di questo genere. Una delle variabili di risposta era una misura dell'importanza data all'amicizia. I risultati sono riassunti nella seguente tabella:

Gruppo	Trattamento	n	media	d.s.
1	Volontariato	57	105.32	14.68
2	Nessuna attività	17	96.82	14.26

```
> test.t2ci(media1=105.32,ds1=14.68,n1=57,
            media2=96.82,ds2=14.26,n2=17)

Primo   campione: media = 105.32  d.s. = 14.68  n = 57
Secondo campione: media = 96.82  d.s. = 14.26  n = 17

Varianza congiunta = 212.8013  ES = 4.031263
Test t = 2.108520  g.l. = 72  p-value = 0.03846607
Intervallo di confidenza al 95% per la differenza:
   da 0.4638239 a 16.53618

Errore standard della differenza = 3.967665
Test di Welch = 2.142318  g.l. = 26.94  p-value = 0.04136
Intervallo di confidenza al 95% per la differenza:
   da 0.3582273 a 16.64177
```

Il modo in cui strumenti e dispositivi sono progettati influisce sul modo in cui le persone riescono ad utilizzarli. In uno studio è stato chiesto a 25 persone destre di girare una manopola (con la loro mano destra) che muoveva un indicatore. C'erano due strumenti identici, uno per destri (la manopola andava girata in senso orario) ed uno per mancini (la manopola andava girata in senso antiorario). Ciascun soggetto ha usato entrambi gli strumenti secondo un ordine casuale. La tabella che segue riporta i tempi in secondi che ciascun soggetto ha impiegato per girare completamente la manopola. Esiste una differenza significativa fra i tempi medi impiegati per girare le due manopole?

Tabella 16.2 Tempi impiegati per girare completamente una manopola

Soggetto	Manopola per destri	Manopola per mancini	Soggetto	Manopola per destri	Manopola per mancini
1	113	137	14	107	87
2	105	105	15	118	166
3	130	133	16	103	146
4	101	108	17	111	123
5	138	115	18	104	135
6	118	170	19	111	112
7	87	103	20	89	93
8	116	145	21	78	76
9	75	78	22	100	116
10	96	107	23	89	78
11	122	84	24	85	101
12	103	148	25	88	123
13	116	147			

```
> destri <-  c(113,105,130,101,138,
+              118,87,116,75,96,122,
+              103,116,107,118,103,
+              111,104,111,89,78,
+              100,89,85,88)
> mancini <- c(137,105,133,108,115,
+               170,103,145,78,107,
+               84,148,147,87,166,
+               146,123,135,112,93,
+               76,116,78,101,123)
>
> test.t2cd(cbind(destri,mancini))
Primo   campione: media = 104.12  d.s. = 15.79641  n = 25
Secondo campione: media = 117.44  d.s. = 27.26273  n = 25
Differenze      : media = -13.32  d.s. = 22.936  n = 25

Test t = 2.903732  g.l. = 24  p-value = 0.00779151
t critico = 2.063899
I.C. al 95% per la differenza: da -22.78751 a -3.852485
```

Esempio 3.2 In una ricerca volta a valutare quali condizioni aiutino le persone sovrappeso a condurre un regolare esercizio fisico, un insieme di persone è stato suddiviso, in modo del tutto casuale, in tre gruppi corrispondenti a tre trattamenti diversi da svolgersi 5 giorni su 7: un solo esercizio fisico in palestra di durata protratta; parecchi esercizi da 10 minuti ciascuno, sempre in palestra; esercizi ripetuti da 10 minuti ciascuno, ma da eseguire a casa impiegando un attrezzo ginnico fornito dai ricercatori. Dopo 6 mesi di trattamento i risultati, in termini di perdita di peso in kg, sono stati i seguenti:

Trattamento	n	media	ds
Unico esercizio lungo	37	10.2	4.2
Molti esercizi brevi	36	9.3	4.5
Esercizi brevi a casa	42	10.2	5.2

All'interno di `rmf` la funzione `OneWay` permette di eseguire l'ANOVA a partire da queste informazioni. I suoi argomenti sono rappresentati da tre vettori: il primo contiene le medie, il secondo le deviazioni standard e il terzo il numero delle osservazioni di ciascun gruppo. Il numero dei gruppi a confronto corrisponde alla lunghezza dei tre vettori, che deve essere la stessa.

```
> m <- c(10.2,9.3,10.2)
> s <- c(4.2,4.5,5.2)
> n <- c(37,36,42)
> OneWay(m,s,n)
            Df Sum Sq Mean Sq F value Pr(>F)
gruppo       2     20   10.02   0.457  0.634
Residuals  112   2452   21.90
```

di “civismo”. A 946 soggetti è stato chiesto di specificare, impiegando una scala da 1 a 10, quanto erano d’accordo con le tre seguenti affermazioni:

- La scienza e la tecnologia stanno rendendo la nostra vita più sana, facile e più confortevole.
- Grazie alla scienza ed alla tecnologia ci saranno maggiori opportunità per la prossima generazione.
- Dipendiamo troppo dalla scienza e non abbastanza dalla fede.

Le risposte individuali a ciascuna delle tre domande (identificate rispettivamente con le etichette A, B e C) sono contenute in forma *long* nel *data-frame* `wvs.long` e in forma *wide* nel *data-frame* `wvs.wide`. In questo ultimo caso le tre colonne del *data-frame* corrispondono alle tre domande

```
> aov.mr(wvs.wide)
Analysis of Variance Table

Greenhouse-Geisser epsilon: 0.6786
Huynh-Feldt epsilon:        0.6792

  Res.Df Df Gen.var.      F num Df den Df   Pr(>F)    G-G Pr    H-F Pr
1    945       1.9964
2    946  1    2.0021 4.6525      2   1890 0.009647 0.020574 0.020544
```

Si scelgono 50 assaggiatori a cui si fanno valutare separatamente due tazze di caffè senza marchi di riconoscimento e poi si chiede loro quale caffè preferi-scono. Una delle due tazze contiene un caffè idro-solubile, l’altra il normale caffè americano. Trentuno assaggiatori preferiscono il caffè americano. Possiamo affermare che la maggioranza delle persone preferisce il caffè americano? Calcoliamo l’i.c. al 90%.

```
> test.prop(31,50,p0=0.5,conf=0.9)
stima di p = 0.62  n = 50
ES = 0.06864401  margine di errore = 0.1129093
I. C. approssimato al 90% per p: da 0.5070907 a 0.7329093
I. C. di Wilson al 90% per p: da 0.5036945 a 0.7239856
I. C. esatto       al 90% per p: da 0.4939593 a 0.7349308

Ipotesi nulla: p = 0.5  ES sotto H0= 0.07071068
Test z approssimato: 1.697056  p-value = 0.08968602
Test esatto binomiale: p-value = 0.1189205
```

La funzione `marascuilo` di `rmf` è in grado di calcolare intervalli di confidenza ed eseguire test di significatività applicando il metodo di Marascuilo. Ne illustriamo di seguito l'impiego sui dati del Titanic.

```
> tbl <- matrix(c(61,111,22,150,85,419),nrow=2)
> colnames(tbl) <- c("Alta","Media","Bassa")
> tmp <- marascuilo(tbl)
> summary(tmp)
                    Diff         SE     chi-2            p
Alta - Media   0.22674419 0.04448808 25.976765 2.286742e-06
Alta - Bassa   0.18600037 0.04011049 21.503630 2.140652e-05
Media - Bassa -0.04074382 0.03044204  1.791332 4.083357e-01

> confint(tmp, level=0.9)
                    Diff       Lower      Upper
Alta - Media   0.22674419  0.13127429 0.32221409
Alta - Bassa   0.18600037  0.09992463 0.27207611
Media - Bassa -0.04074382 -0.10607140 0.02458376
```