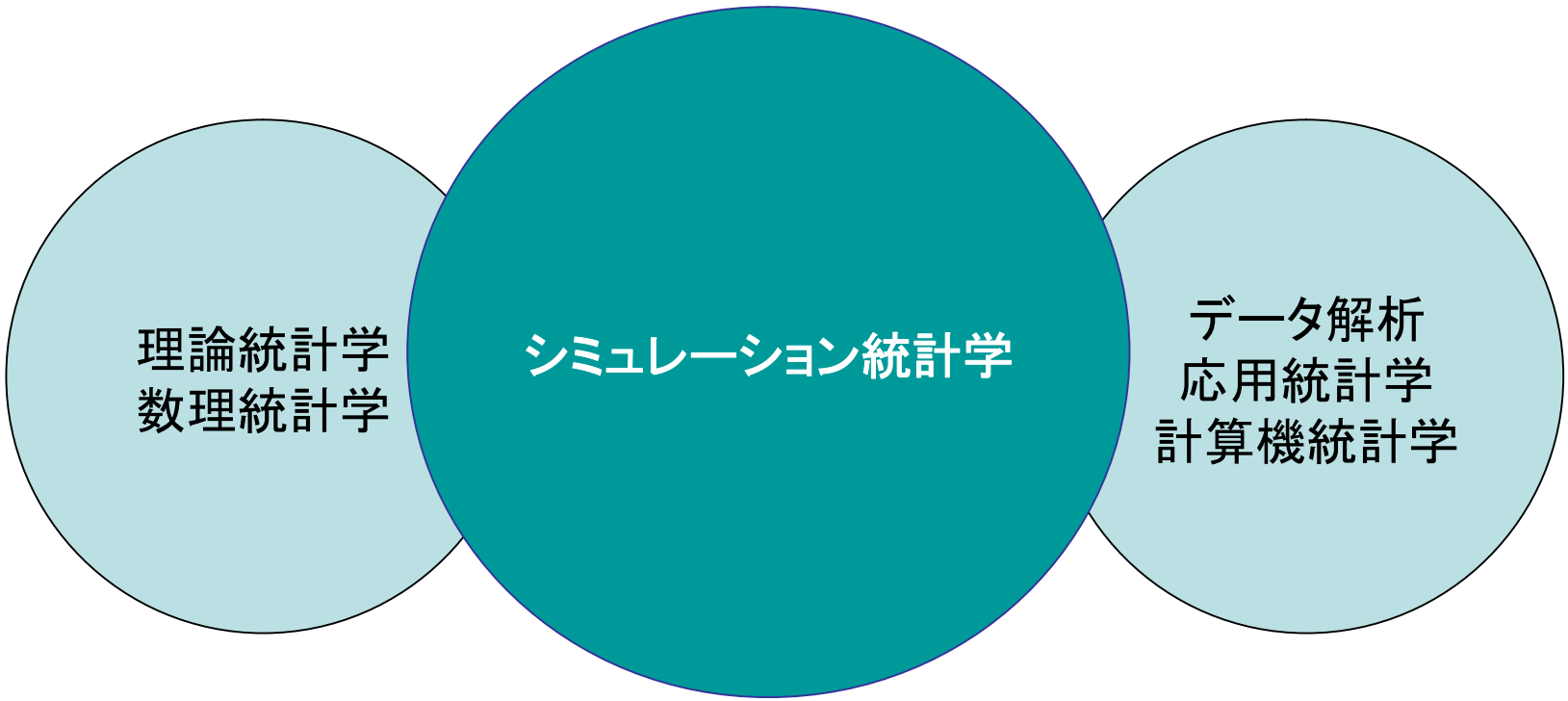
シミュレーション統計学の勧め

岡山大学
垂水共之

統計学

- 理論統計学（数理統計学）
  - 確率論をベースに
  - 分布論
  - 推測統計学

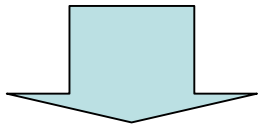
- データ解析（応用統計学）
  - 実データの分析
  - 表計算ソフト
  - 統計ソフト


理論統計学
数理統計学
シミュレーション統計学
データ解析
応用統計学
計算機統計学

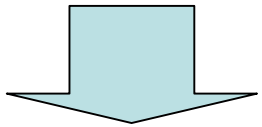
特定研究

- 「情報システムの形成過程と学術情報の組織化」（猪瀬博）昭和51年度～53年度
- A4班「統計プログラム・パッケージの研究」（代表：丘本正）
  - NISAN
  - SALS
  - 行動科学
  - 方法論

- 汎用機のTSS処理によるシミュレーション教育の試み（千葉大学・田栗・現在大学入試センター）



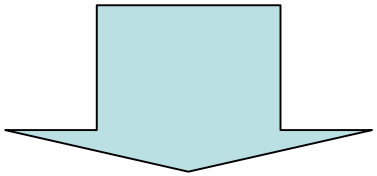
- パソコン（NEC PC－8001）への焼き直し
  - スピードが遅く挫折



- パソコンの低価格化と高性能化

シミュレーション統計学

- パソコンの低価格化と高性能化
  - 本体は3～4万円で充分の性能
- 統計解析用ソフト
  - XLisp－Stat
  - R/S-PLUS



R/S-PLUSによる統計解析入門
（2006年、共立出版）

R/S-PLUSによる統計解析入門

- 乱数によるシミュレーション
  - 円周率
- 標本分布のシミュレーション
  - 標準正規分から自由度1のχ２乗分布
  - χ２乗分布の再生性
  - 標準正規分とχ２乗分布の比としての t分布
  - ２つのχ２乗分布の比としてのF分布
- 中心極限定理
- 正規分布の母平均の区間推定
  - 信頼度
- 正規分布の母平均の検定
  - 有意水準
  - 検出力

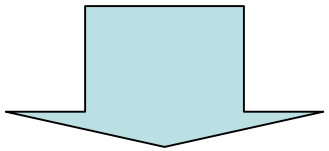
S-PLUSの環境

- Windows版
  - S-PLUS 6.0 日本語版 Release 4
  - コマンド（CGI）での利用
- Linux版
  - Version 6.0 Release 1

中心極限定理

中心極限定理
Central Limit Theorem

- X1, X2, ..., Xn は同じ分布 F(x) に従い独立
- $E(X)=\mu$ 平均存在
- $Var(X)=\sigma^2$ 分散存在



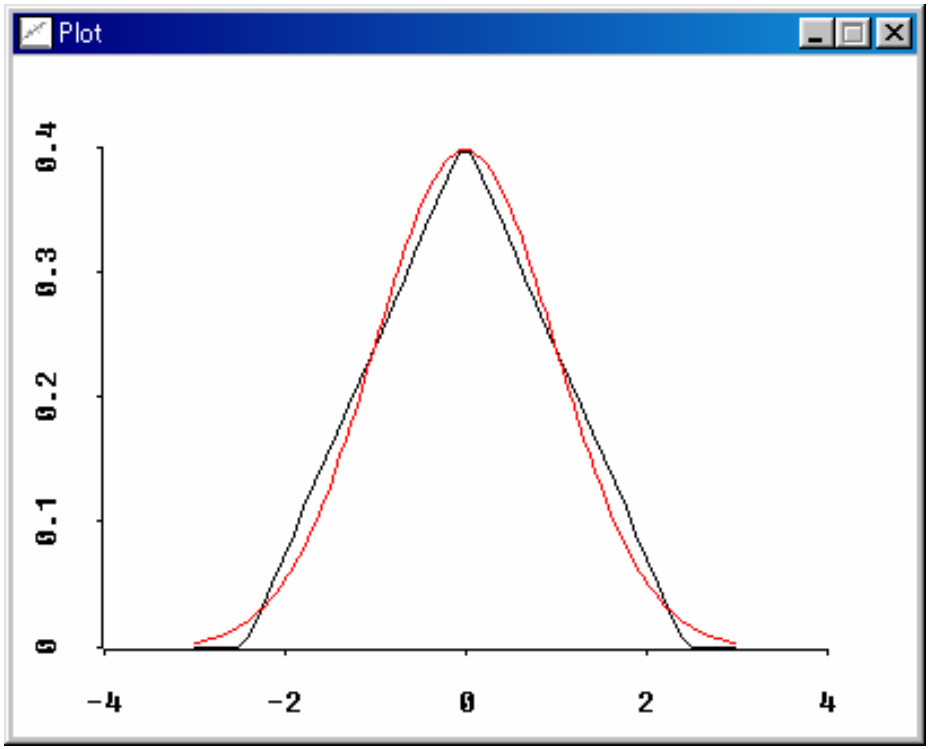
- $Xbar=\Sigma X_i/n$ の分布は $n\to\infty$ のとき
  $N(\mu, \sigma^2/n)$ に収束

- 実用的には $n \geqq 25$

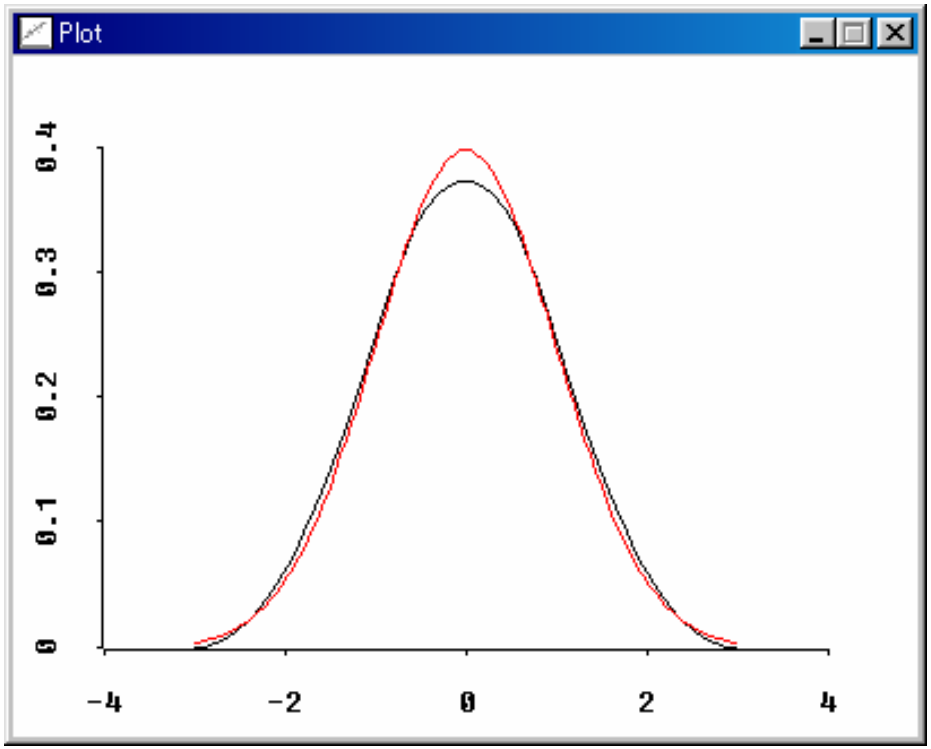
一様分布の場合

- X が区間 [0, 1] の一様分布
  E(X)=1/2, Var(X)=1/12

n=2



n=3



$X_i \sim \quad U[0,1]$

$Z = X_1$ のpdf

$$u_1(x) = \begin{cases} 1 & 0 < x < 1 \\ 0 & \text{その他} \end{cases}$$

$Z = X_1 + X_2$ のpdf

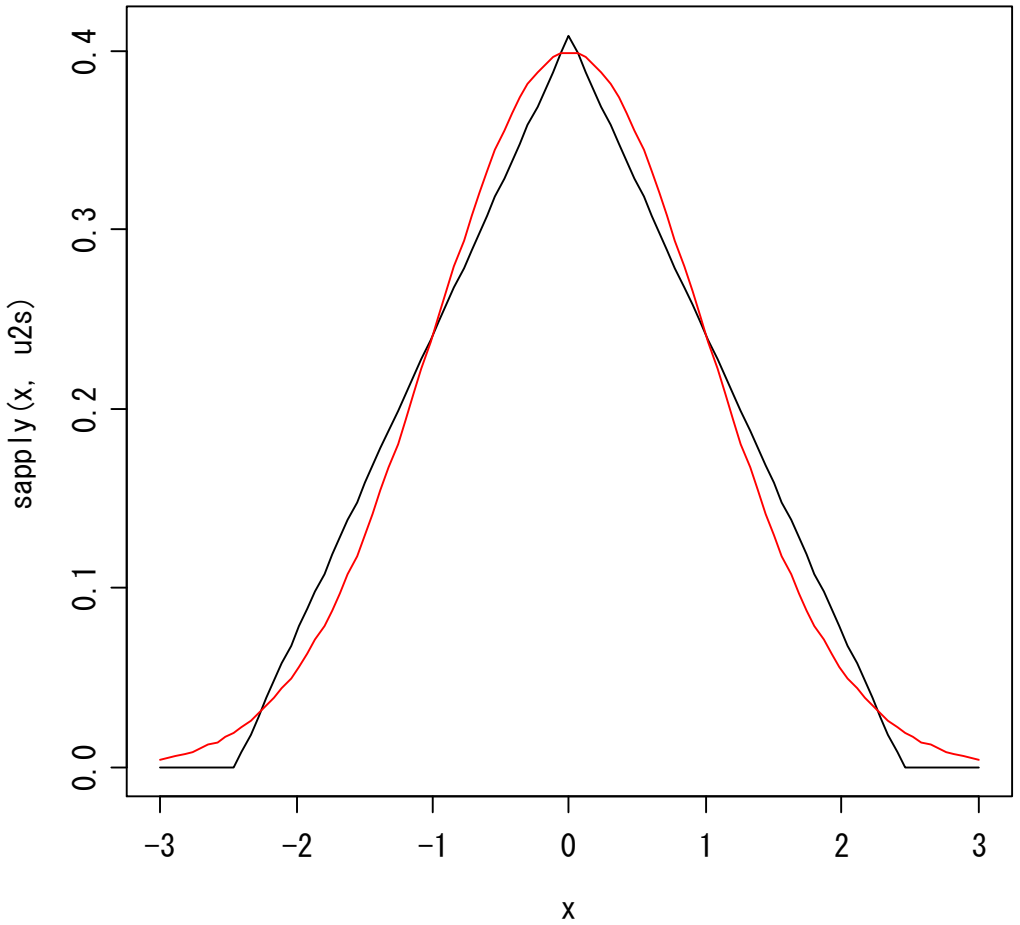
$$u_2(x) = \begin{cases} z & 0 < z < 1 \\ 2 - z & 1 < z < 2 \\ 0 & \text{その他} \end{cases}$$

$Z = X_1 + X_2 + X_3$ のpdf

$$u_3(x) = \begin{cases} z^2 & 0 < z < 1 \\ -z^2 + 3z - 3/2 & 1 < z < 2 \\ (z-3)^2/2 & 2 < z < 3 \\ 0 & \text{その他} \end{cases}$$

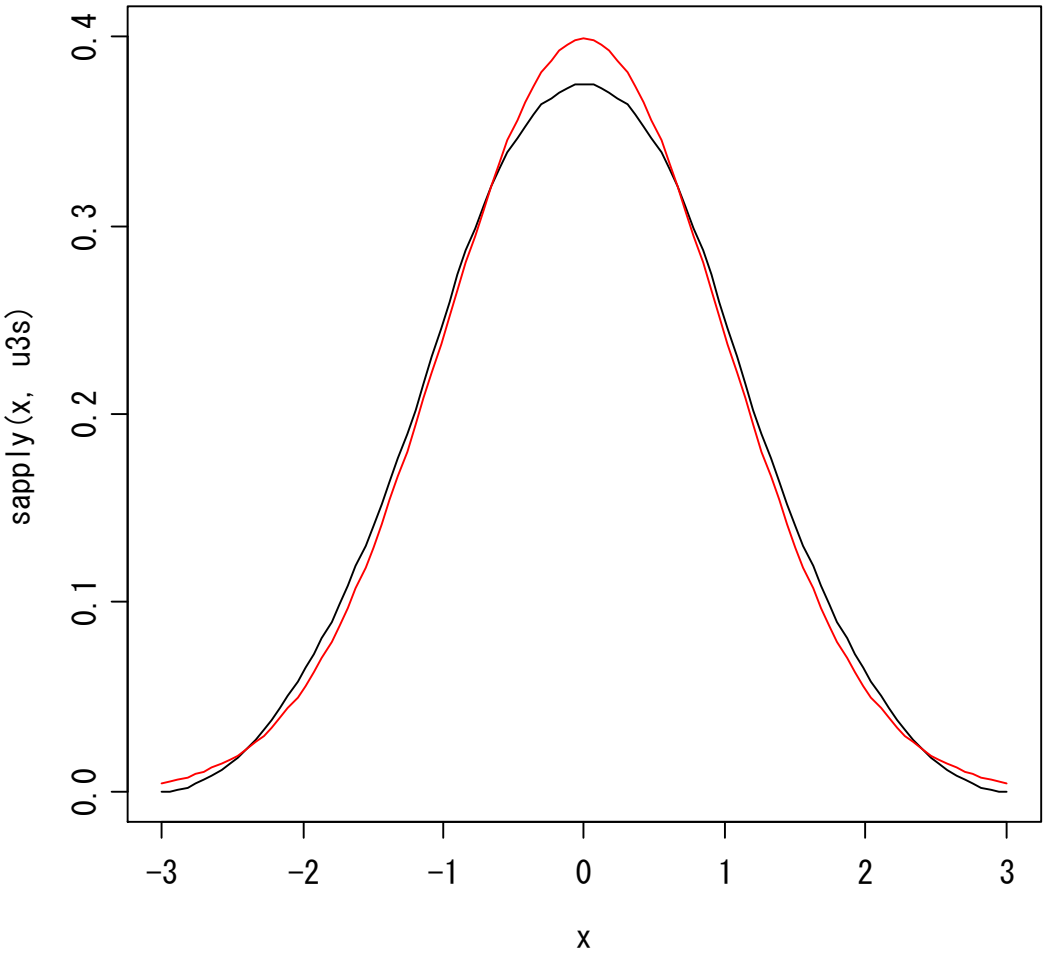
二個の一様乱数の和

```
> u2 <- function(z)
+   switch(length(which(c((z >= 0), (z >= 1), (z > 2)))) + 1, 0, z, 2 - z, 0)
> u2s <- function(z)
+   u2(z / sqrt(6) + 1) / sqrt(6)

> x <- seq(-3, 3, length = 101)
> plot(x, sapply(x, u2s), type = "l")
> curve(dnorm, -3, 3, add = T, col = 2)
```



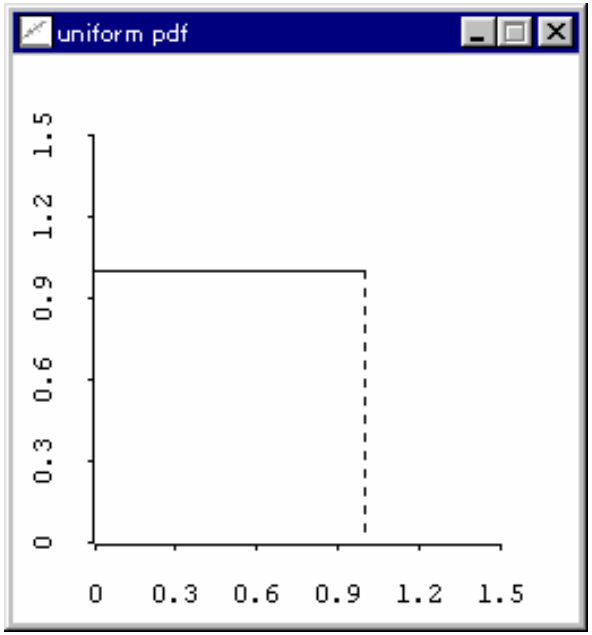
三個の一様乱数の和

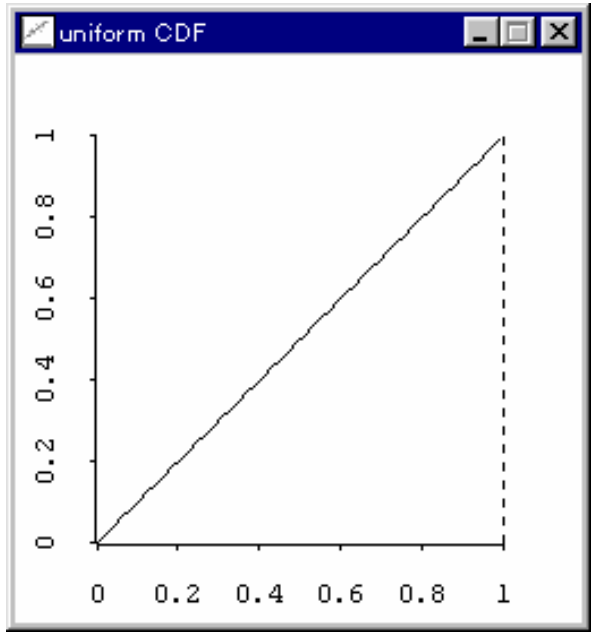
```
> u3 <- function(z)
+   switch(length(which(c((z >= 0), (z >= 1), (z > 2), (z > 3)))) + 1,
+   0, z^2 / 2, -z^2 + 3 * z - 3 / 2, (z - 3)^2 / 2, 0)
> u3s <- function(z)
+   u3(1.5 + (z * 0.5))* 0.5
> x <- seq(-3, 3, length = 101)
> plot(x, sapply(x, u3s), type = "l",
+    ylim = c(0, 0.4))
> curve(dnorm, -3, 3, add = T, col = 2)
```



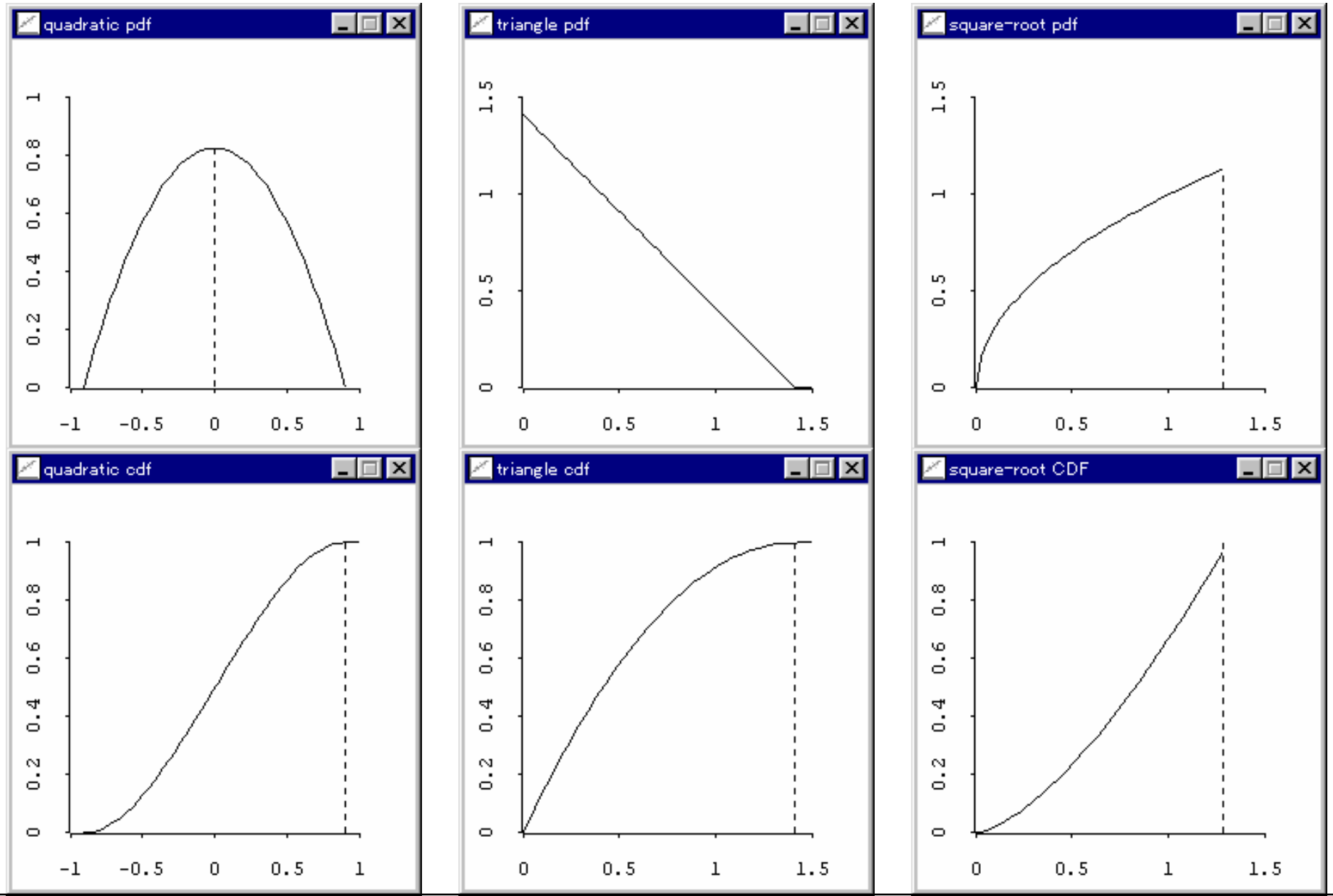
種々の分布

- 一様分布
- 二次分布
- 三角分布
- 平方根分布
- 二項分布
- …





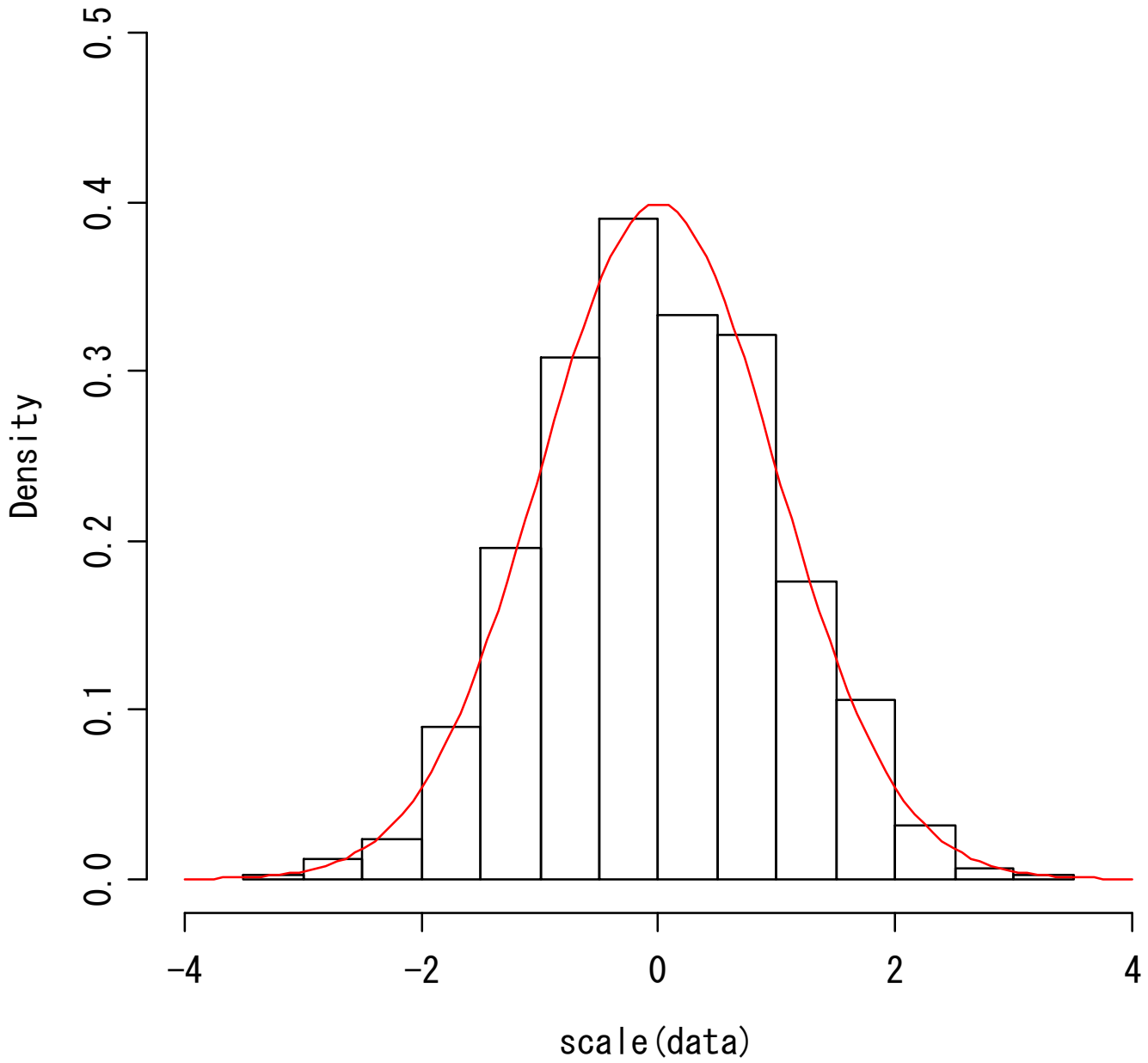
種々の分布のpdf, cdf



シミュレーション(一様分布)

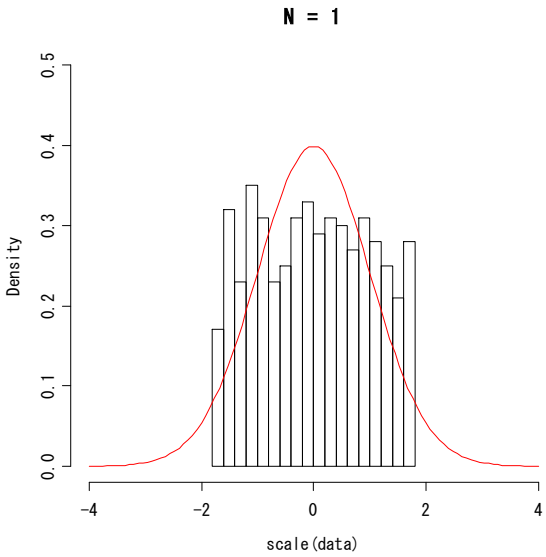
```
> cltplot <- function(data, n){
+   hist(scale(data), nclass = 20, xlim = c(-4, 4),
+   ylim = c(0, 0.5), freq = F, main = paste("N =" , n))
+   curve(dnorm, -4, 4, col = 2, add = T)
+ }

> clturand <- function(n, nsim){
+   if(n == 1)
+     result <- apply(matrix(sapply(rep(n, nsim), runif), n, nsim), 2, mean)
+   else
+     result <- apply(sapply(rep(n, nsim), runif), 2, mean)
+   cltplot(result, n)
+ }



> clturand(5,1000)
```

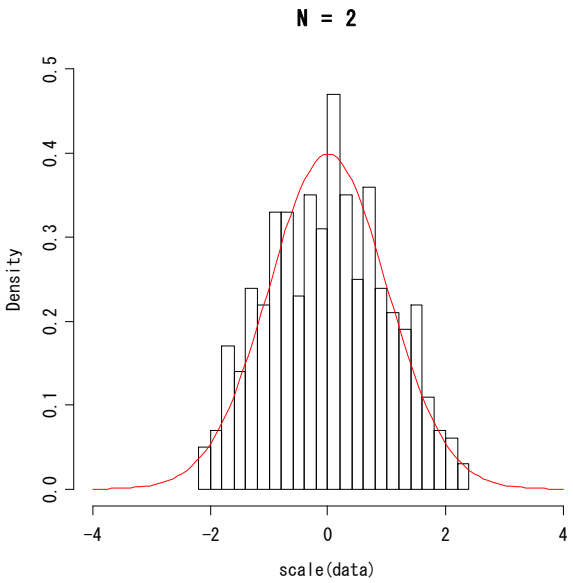
clturand(5,1000)

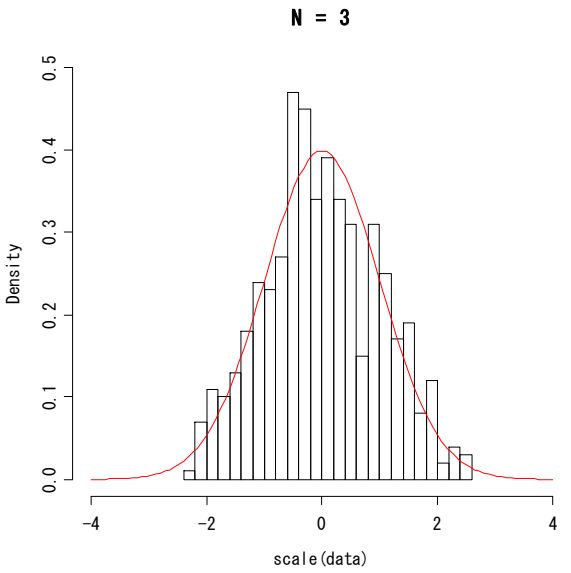


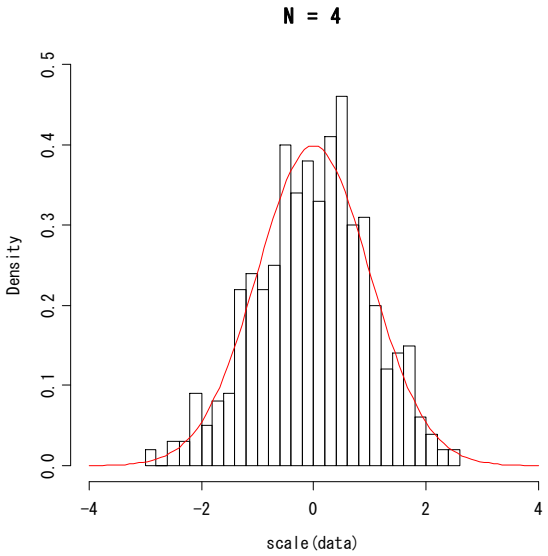
cltunif(5)　データ数を１から5まで

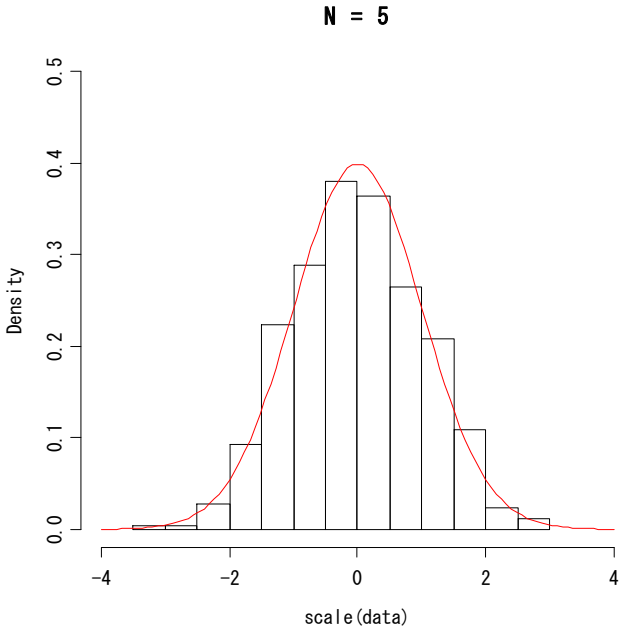
```
> cltunif <- function(nmax){
+   for( i in 1:nmax){
+ if (options()$device == "X11")
+       X11()
+     if (options()$device == "windows")
+       win.graph()
+     clturand(i, nsim)
+   }
+ }
>
> ####
> nsim <- 500
> #   標準のシミュレーション回数
> cltunif(5)
```


N = 1
Density
0.0
0.1
0.2
0.3
0.4
0.5
−4
−2
0
2
4
scale(data)


N = 2
Density
0.0
0.1
0.2
0.3
0.4
0.5
−4
−2
0
2
4
scale(data)


N = 3
Density
0.0
0.1
0.2
0.3
0.4
0.5
−4
−2
0
2
4
scale(data)

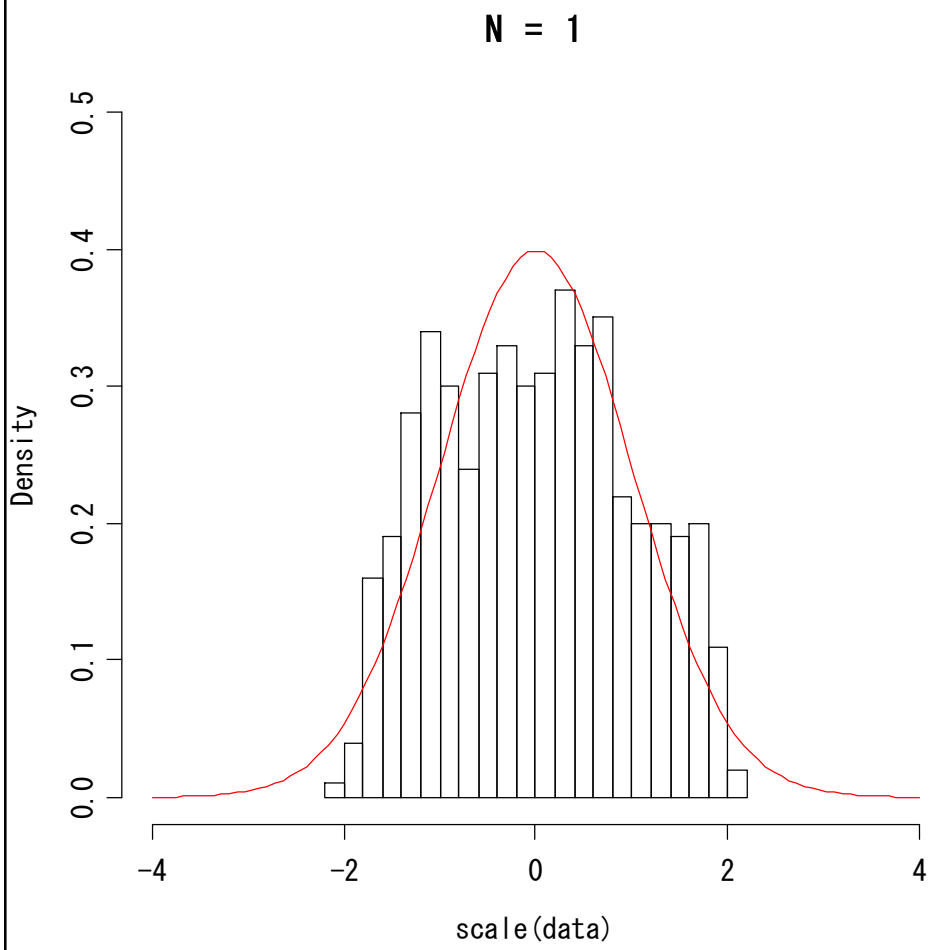

N = 4
Density
0.0
0.1
0.2
0.3
0.4
0.5
−4
−2
0
2
4
scale(data)

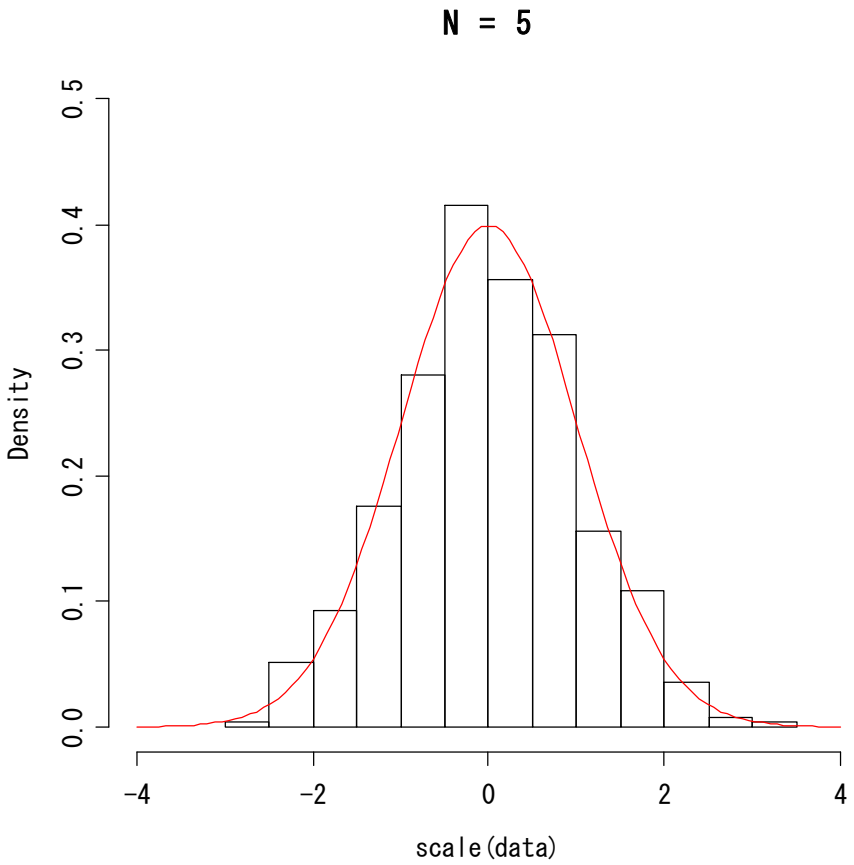

N = 5
Density
0.0
0.1
0.2
0.3
0.4
0.5
−4
−2
0
2
4
scale(data)

シミュレーション
乱数（分布）を指定

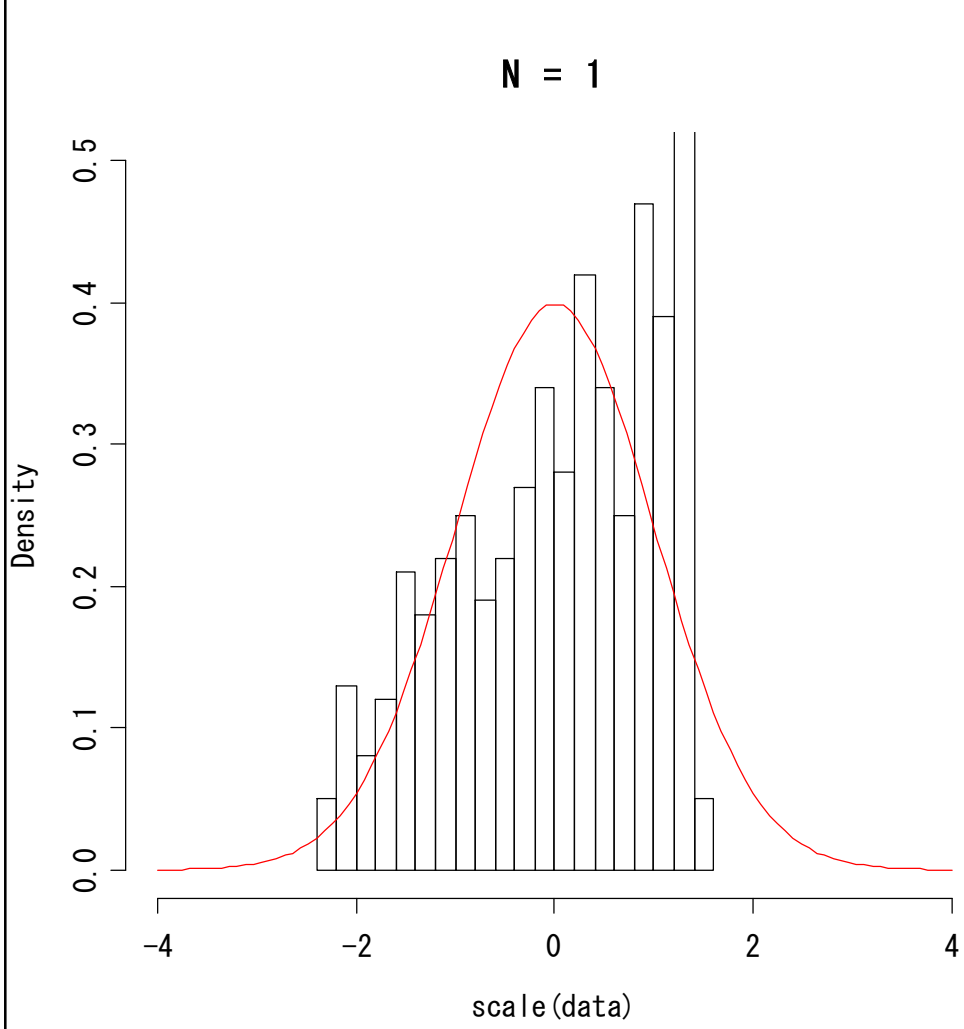
```
cltrand <- function(n, nsim, rdist){
  if(n == 1)
    result <- apply(matrix(sapply(rep(n, nsim), rdist), n, nsim), 2, mean)
  else
    result <- apply(sapply(rep(n, nsim), rdist), 2, mean)
  cltplot(result, n)
}

clt <- function(nmax, rdist){
  for(i in 1:nmax){
    if (options()$device == "X11")
      X11()
    if (options()$device == "windows")
      win.graph()
    cltrand(i, nsim, rdist)
  }
}
```

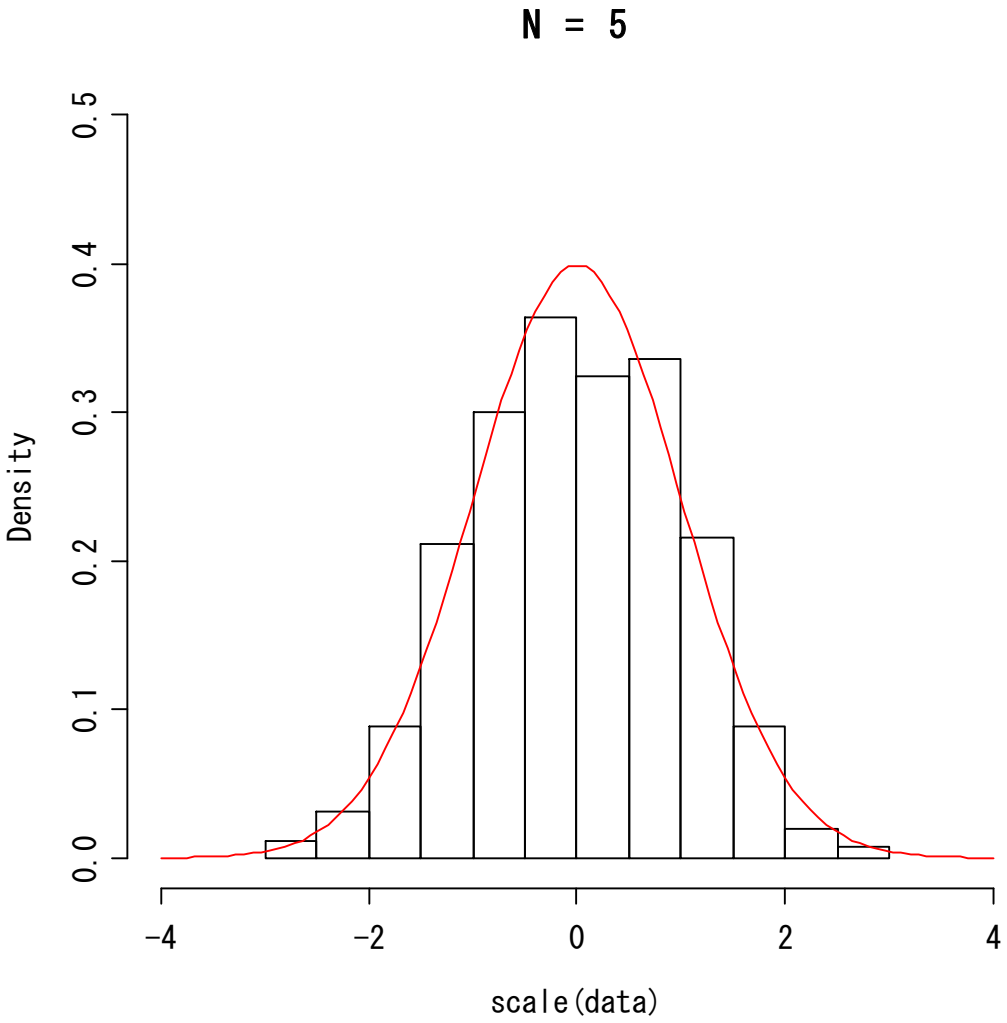
> clt(5,rquad)


N = 1
Density
0.0
0.1
0.2
0.3
0.4
0.5
−4
−2
0
2
4
scale(data)


N = 5
Density
0.0
0.1
0.2
0.3
0.4
0.5
−4
−2
0
2
4
scale(data)

rsqrt(5,500)


N = 1
Density
0.0
0.1
0.2
0.3
0.4
0.5
−4
−2
0
2
4
scale(data)


N = 5
Density
0.0
0.1
0.2
0.3
0.4
0.5
−4
−2
0
2
4
scale(data)

シミュレーション(その他の分布)

- 二次分布　　rquad
- 三角分布　　rtriangular
- 平方根分布　rsqrt
- 二項分布　　rbinom

...

> brinom10.01<-function(x) rbinom(n,10,0.01)
> clt(15, clt-binomial)

推定・検定

区間推定の信頼度 1-α

$$\Pr(\hat{\theta}_L(X_1, X_2, \cdots, X_n) < \theta < \hat{\theta}_U(X_1, X_2, \cdots, X_n)) < 1-\alpha$$

信頼度 $1-\alpha$ の信頼区間

$$(\hat{\theta}_L(X_1, X_2, \cdots, X_n), \hat{\theta}_U(X_1, X_2, \cdots, X_n))$$

- 信頼度 1-α=0.95 とは
  - 求めた区間に真の値が入る確率が95%
  - 100回に5回ほどは外れても仕方がない

母平均の区間推定（分散既知）

- 正規母集団
  - 母分散：1（既知）
  - 母平均：μ（未知）

$$\bar{x} \pm k_{\alpha/2}\sqrt{\frac{\sigma^2}{n}}$$

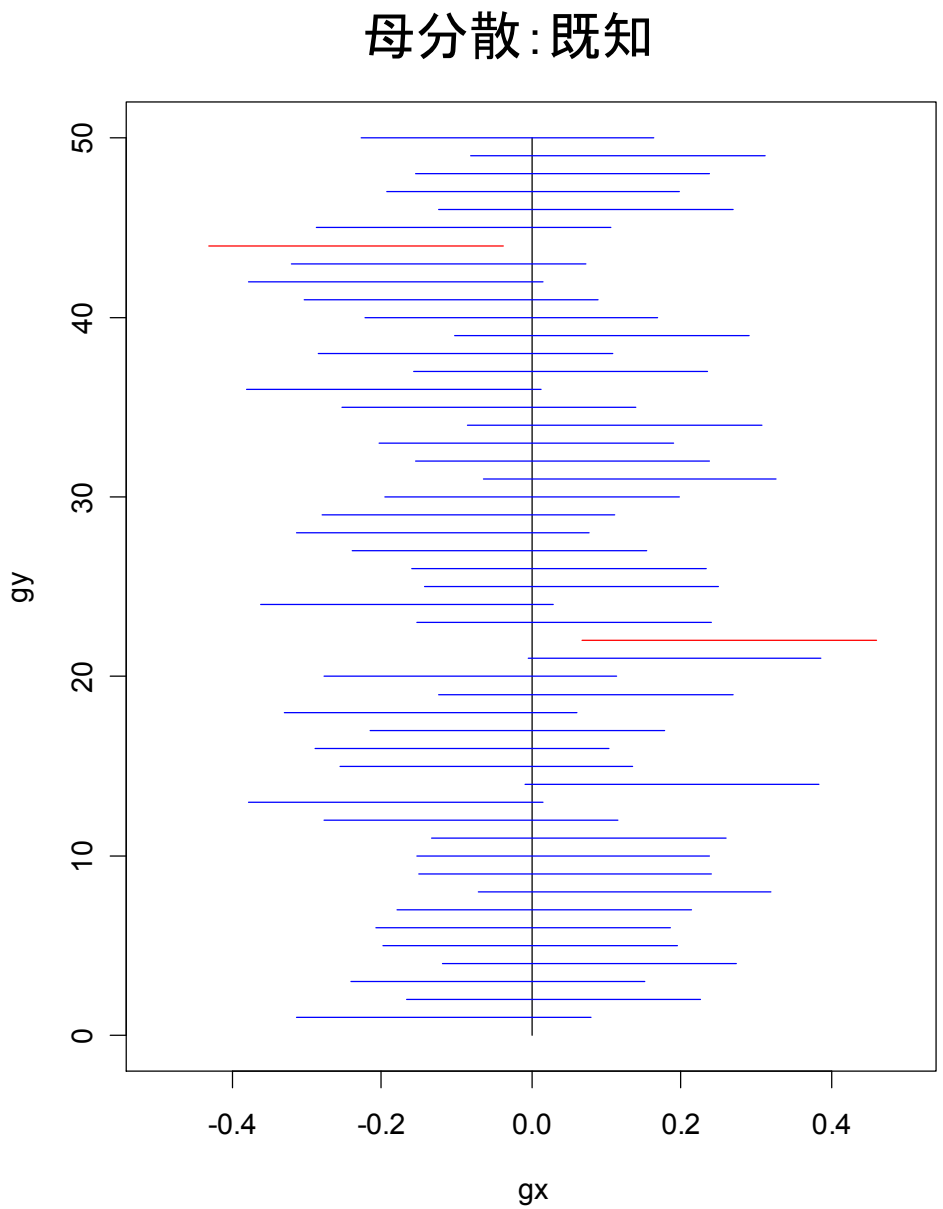
  - データは
    rnorm()
    N(0,1) からのデータ
  - 真の μ =0

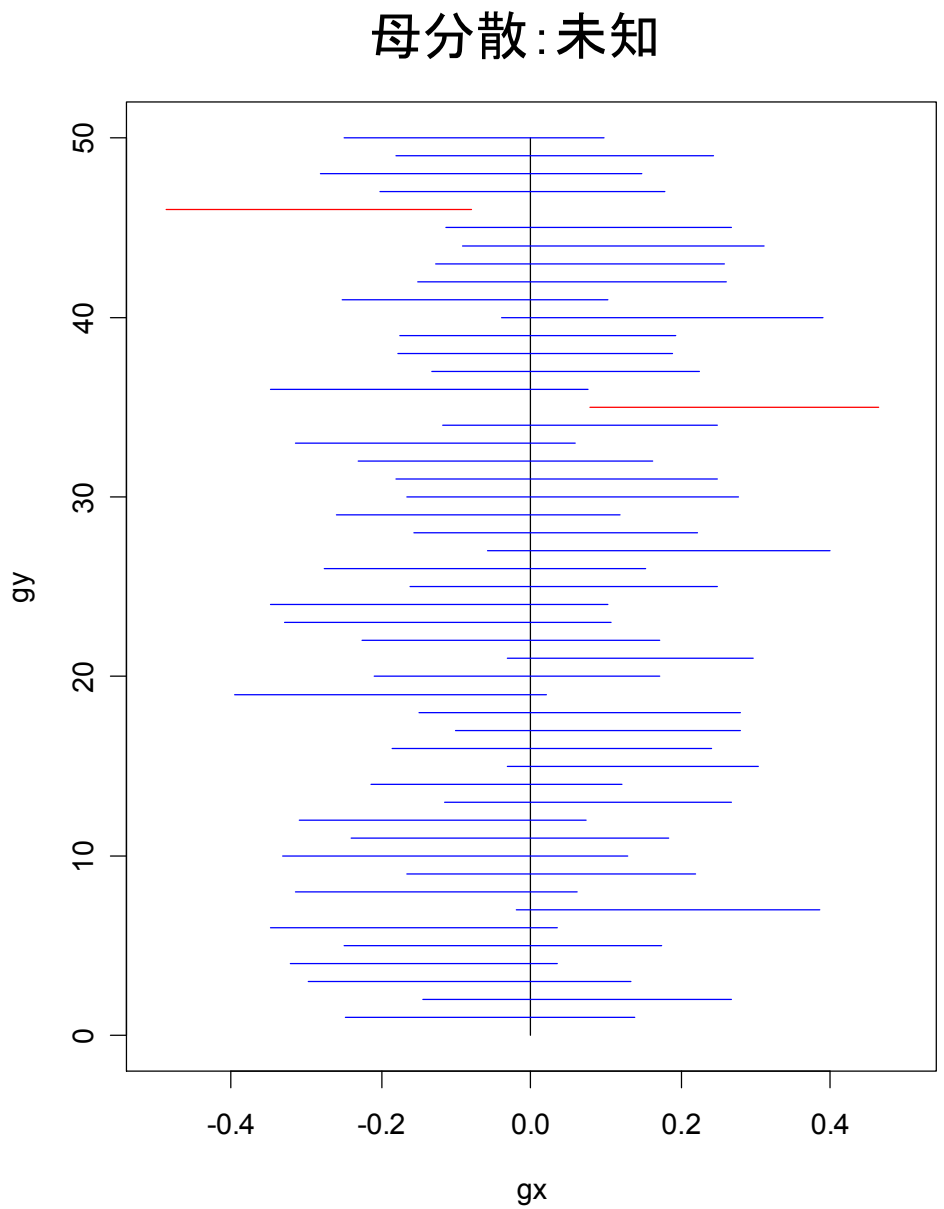
```
> confinterval <- function(x, conf, sigma2)
+ {
+   n <- length(x)
+   alpha <- 1 - conf
+   k <- qnorm(1 - alpha / 2)
+   barx <- mean(x)
+   mL <- barx - k * sqrt(sigma2 / n)
+   mU <- barx + k * sqrt(sigma2 / n)
+   c(mL, mU)
+ }
> confinterval(rnorm(10), 0.95, 1)
[1] -0.547104  0.692486
> confinterval(rnorm(10), 0.95, 1)
[1] -0.6639466  0.5756434
> confinterval(rnorm(10), 0.95, 1)
[1] -0.8003083  0.4392818
> confinterval(rnorm(10), 0.95, 1)
[1] 0.01251542 1.25210548
```

100回中、何回程度間違える？

```
> confintervalsim <- function(nsim, n, conf){
+   result <- c()
+   for (i in 1:nsim){
+     result <- rbind(result, confinterval(rnorm(n), conf, 1))
+   }
+   result
+ }

> r <- confintervalsim(100, 10, 0.95)
> r[apply(r, 1, prod) > 0, ]
          [,1]         [,2]
[1,] -1.2425114 -0.002921355
[2,]  0.3166913  1.556281376
[3,]  0.1919863  1.431576354

> which(apply(r, 1, prod) > 0)
[1] 13 31 61
```

母分散、既知と未知


母分散:既知
gy
gx
0
10
20
30
40
50
-0.4
-0.2
0.0
0.2
0.4


母分散:未知
gy
gx
0
10
20
30
40
50
-0.4
-0.2
0.0
0.2
0.4

検定の有意水準

- 有意水準 α=5% とは
  - 帰無仮説が正しいのに、誤って帰無仮説を棄却する、「検定の第一種の誤り」の確率
  - 誤った判定を下す許容範囲
  - 100回中5回程度は間違えても仕方がない！

母平均の検定

$$H_0 : \mu = 0$$

$$H_1 : \mu \neq 0$$

$$\sigma^2 = 1 \text{ (既知)}$$

母集団: 標準正規分布

標本数: n

有意水準: α

```
> ntest <- function(n, alpha)
+ {
+   k <- qnorm(1 - alpha / 2)
+   x <- rnorm(n)
+   barx <- mean(x)
+   z <- barx / sqrt(1 / n)
+   if (abs(z) > k)
+     T
+   else
+     F
+ }

> ntest(10,0.05)
[1] F
> ntest(10,0.05)
[1] T
```

100回中、何回程度間違える？

```
> ntestsim <- function(nsim, n, alpha)
+ {
+   r <-c()
+   for (i in 1:nsim)
+     r <-c(r, ntest(n, alpha))
+   r
+ }

> rslt <- ntestsim(100,10,0.05)
> rslt
  [1] F F T F T F F F F F F F F F F F F F F F F F T F F F F F F F F F F F F F F F F F F F F
 [43] F F F T F F F F F F F F F F F F F F F F F F F F F T F F F F F F F F F F F F F F F F
 [85] F F F F F F F F F F T F F F F F

> which(rslt)
[1]  3  5 24 46 68 95
```

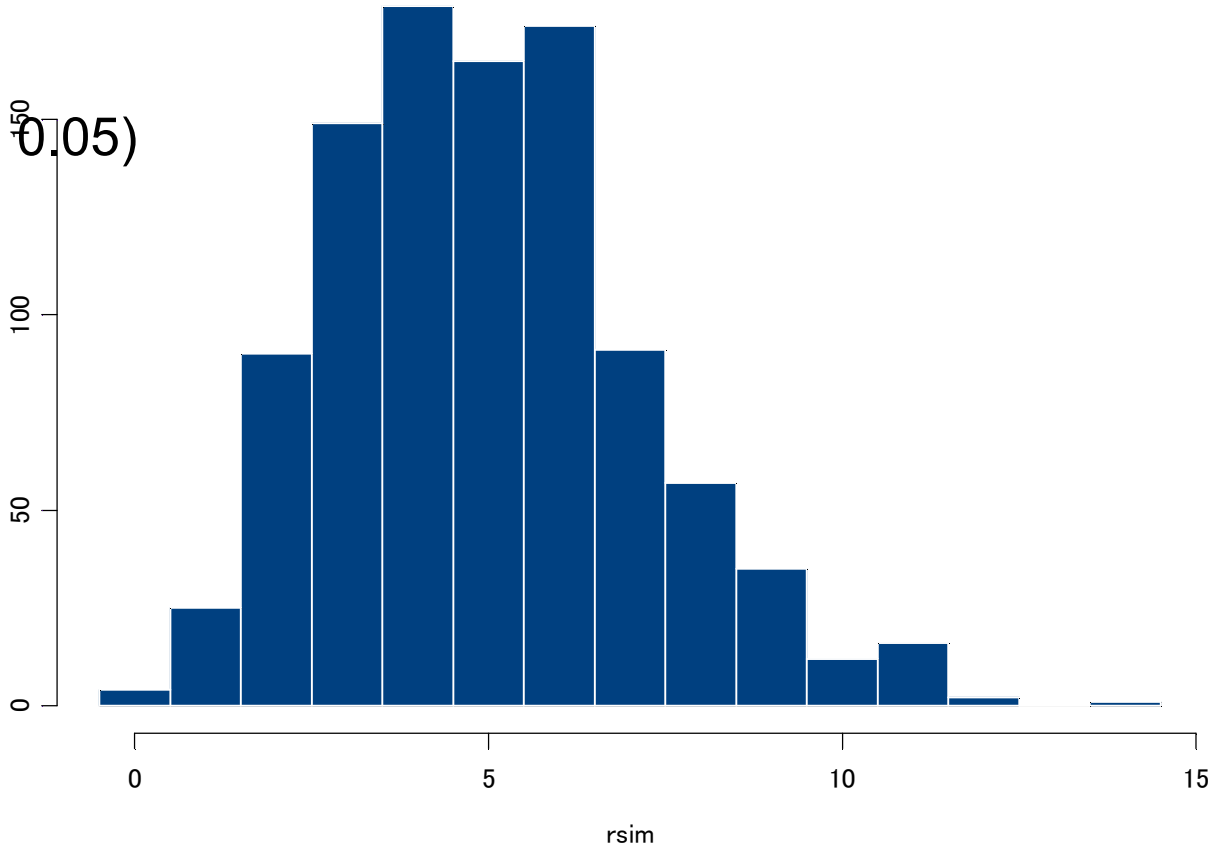
6回

つづき

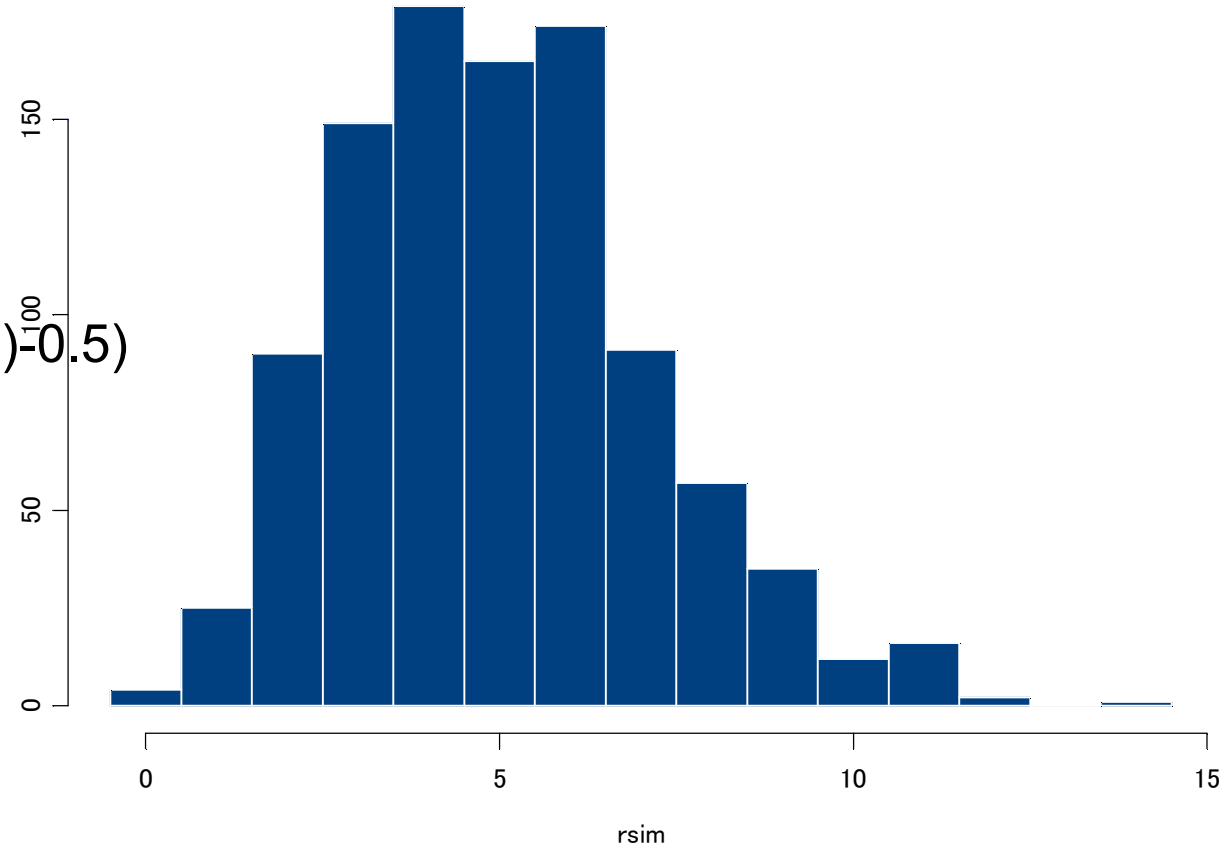
```
> which(ntestsim(100, 10, 0.05))
[1]  12  38  52  53 100                    5回

> which(ntestsim(100, 10, 0.05))
[1]  4 24 30 77 78 81 90                    7回

> which(ntestsim(100, 10, 0.05))
[1]  2 17 53 86 96 98                       6回

> which(ntestsim(100, 10, 0.05))
[1]  3  4 18 37 80                          5回

> which(ntestsim(100, 10, 0.05))
[1] 23 36 60 94                             4回
```

何回ぐらい間違えることが多い？

```
> histsig <- function(nrepeat, nsim, n, alpha)
+ {
+   nsig <- c()
+   for (i in 1:nrepeat)
+     nsig <- c(nsig, length(which(ntestsim(100, 10, 0.05))))
+   nsig
+ }

> rsim <- histsig(1000, 100, 10, 0.05)

> min(rsim)
[1] 0

> max(rsim)
[1] 14
```

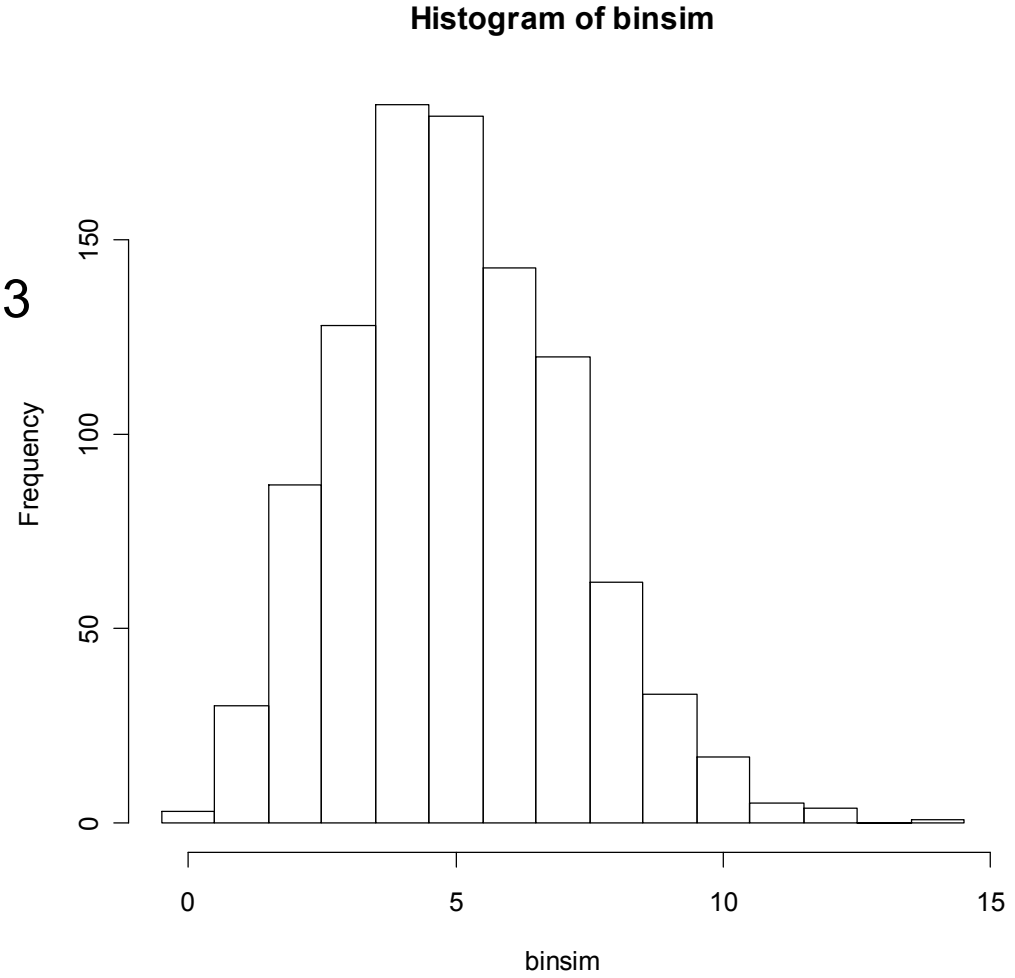


```
> count<-hist(rsim,plot=F,breaks=(0:15)-0.5)
> count
$breaks:
 [1] -0.5  0.5  1.5  2.5  3.5  4.5  5.5  6.5  7.5  8.5  9.5 10.5 11.5 12.5
[15] 13.5 14.5

$counts:
 [1]   4  25  90 149 179 165
 [7] 174  91  57  35  12  16
 [13]  2   0   1

> hist(rsim,plot=T,breaks=(0:15)-0.5)

> mean(rsim)
[1] 4.979
```



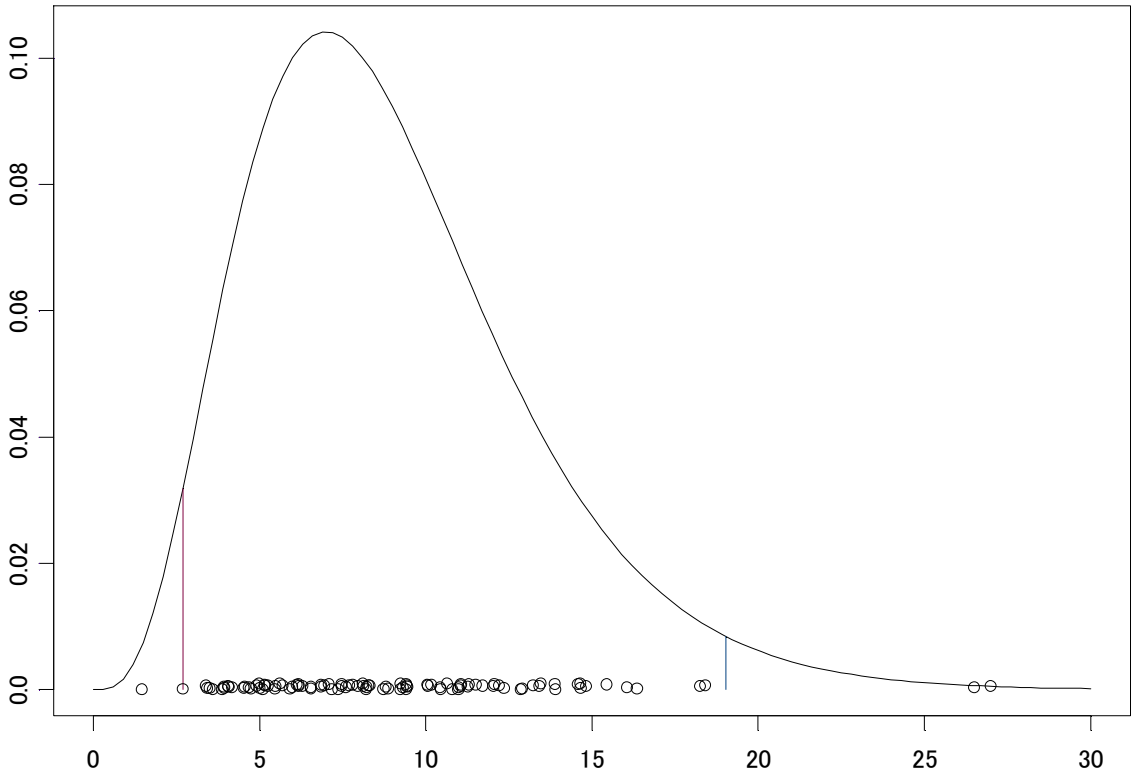
二項分布 Bi(100,0.05)

```
> binsim<-rbinom(1000,100,0.05)
> hist(binsim)
> min(binsim)
[1] 0
> max(binsim)
[1] 14
> hist(binsim,breaks=(0:15)-0.5)
> hist(binsim,breaks=(0:15)-0.5)$counts
 [1]   3  30  87 128 185 182 143 120  62  33
[10]  17   5   4   0   1
> mean(binsim)
[1] 5.016
```



母分散の検定

```
> chiL <- qchisq(0.025, 9); chiL
[1] 2.700389
> chiU <- qchisq(0.975, 9); chiU
[1] 19.02277
> curve(chisqdens9,0,30)
> lines(c(chiU, chiU), c(0, dchisq(chiU, 9)), col = 2)
> lines(c(chiL, chiL), c(0, dchisq(chiL, 9)), col = 3)
> x <-matrix(rnorm(1000), 10, 100)
> tmp1 <- apply(x, 2, sd)
> ssq <-tmp1 * tmp1 * 9
> chi0 <- ssq / 1
> points(ssq, runif(100) * 0.001)
> for (i in chi0){
+   if(i < chiL || i > chiU)
+     print(i)
+ }
[1] 26.50768
[1] 26.99475
[1] 1.471944
[1] 2.700377
```



検出力

- 対立仮説が正しいときに、それを検出する確率

$$H_0 : \mu = \mu_0$$
$$H_1 : \mu = \mu_1$$

$$L(\mu_1) = \Pr(H_1 \mid \mu = \mu_1)$$

母平均の場合

$H_0 : \mu = 0$

$H_1 : \mu = \mu_1$

$N(\mu_1, 1)$のときに、H_0を棄却

```
> test.power.1 <- function(n, alpha, mu1)
+ {
+   k <- qnorm(1 - alpha / 2)
+   x <- rnorm(n, mean = mu1)
+   barx <- mean(x)
+   z <- barx / sqrt(1 / n)
+   if (abs(z) > k)
+     T
+   else
+     F
+ }
```

```
> test.power.1(10, 0.05, 0)
[1] F
```

標本数 n=10　有意水準α=0.05
$\mu_1 = 0$　のとき、検定は受容(F)

```
> test.power.1(10, 0.05, 0.5)
[1] T
```

標本数 n=10　有意水準α=0.05
$\mu_1 = 0.5$のとき、検定は棄却(T)

nsim回繰り返したとき、どの程度棄却するか？

```
> test.power <- function(nsim, n, alpha, mu1)
+ {
+   r <- c()
+   for (i in 1:nsim)
+     r <-c(r, test.power.1(n, alpha, mu1))
+   length(which(r)) / nsim
+ }

> test.power(1000, 10, 0.05, 0)
[1] 0.066
> test.power(1000, 10, 0.05, 0.1)
[1] 0.062
```

```
> mu1 <- seq(-2, 2, 0.1)
```

μ_1 の値を-2から2まで0.1刻みで

```
> calc.power <- function(nsim, n, alpha, mu)  {
+   rslt <- c()
+   for(i in mu)
+     rslt <-c(rslt, test.power(nsim, n, alpha, i))
+   rslt
+ }
```

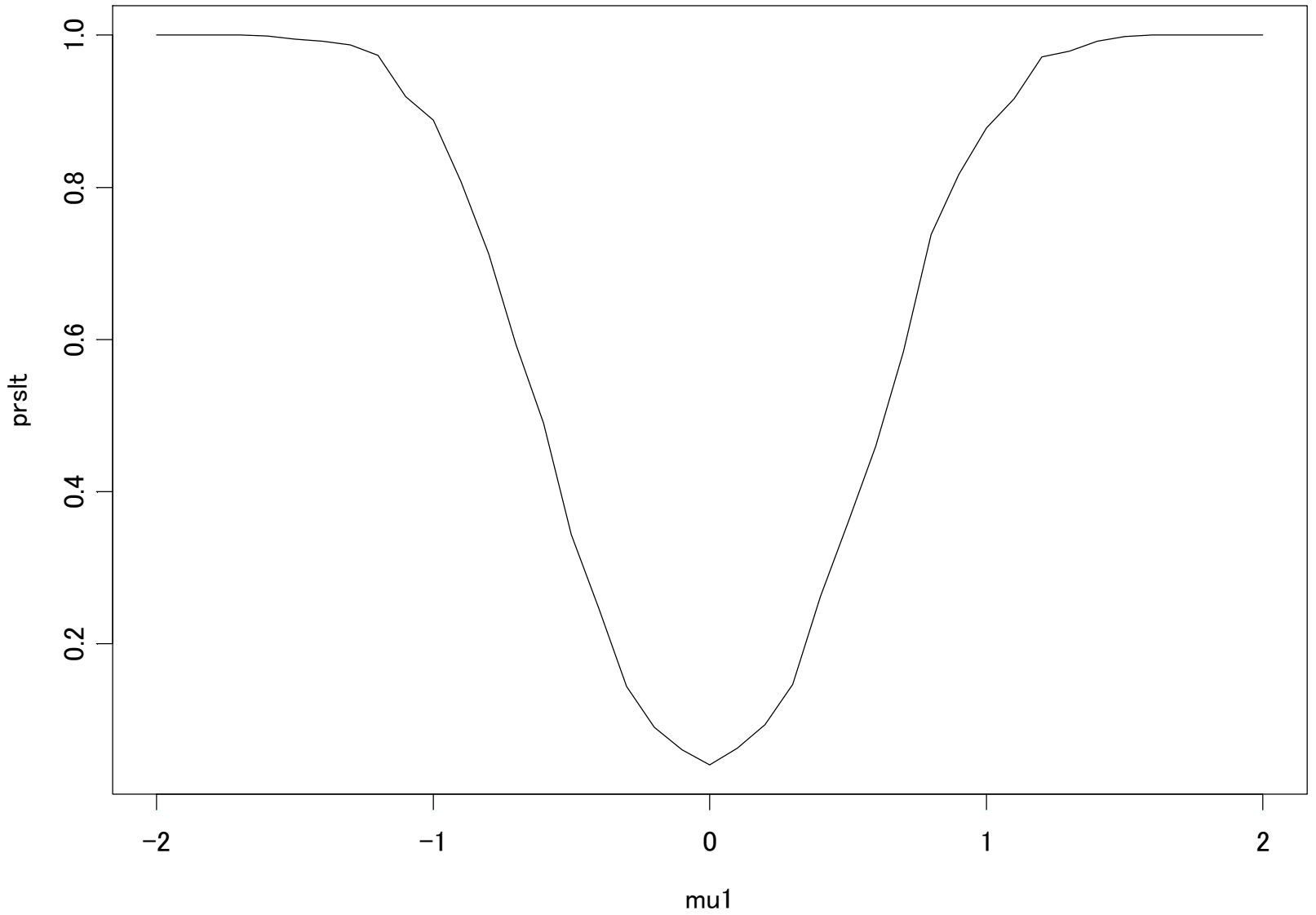
標本数n=10,有意水準 α =0.05

```
> prslt <- calc.power(1000, 10, 0.05, mu1)
```

で各 μ_1 に対して1000回の実験

```
> prslt
 [1] 1.000 1.000 1.000 1.000 0.999 0.995 0.992 0.987 0.973 0.919 0.888 0.807 0.712
[14] 0.594 0.490 0.344 0.246 0.144 0.090 0.061 0.041 0.063 0.094 0.147 0.262 0.360
[27] 0.460 0.584 0.738 0.817 0.878 0.916 0.971 0.979 0.992 0.998 1.000 1.000 1.000
[40] 1.000 1.000
> plot(mu1, prslt, type = "l")
```

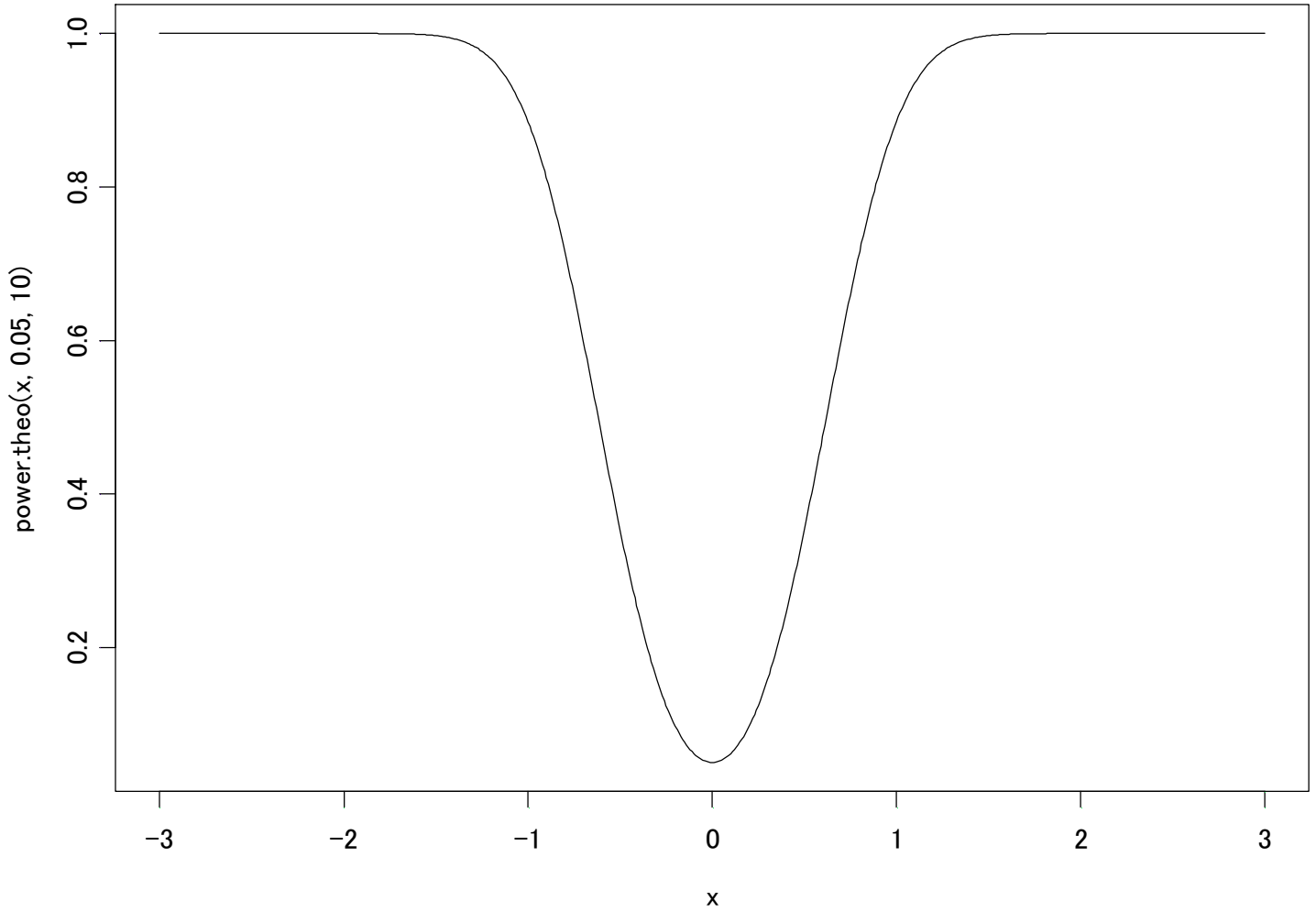
シミュレーションによる検出力関数



理論的には

```
> power.theo <- function(mu, alpha, n)  {
+   qa <- qnorm(1 - alpha / 2)
+   tmp <- (mu - 0) / sqrt(1 / n)
+   err2 <- pnorm(qa - tmp) - pnorm(- qa - tmp)
+   1 - err2
+ }
```

```
> x <- seq(-3, 3, 0.01)
> plot(x, power.theo(x, 0.05, 10), type = "l")
```



標本相関係数の分布

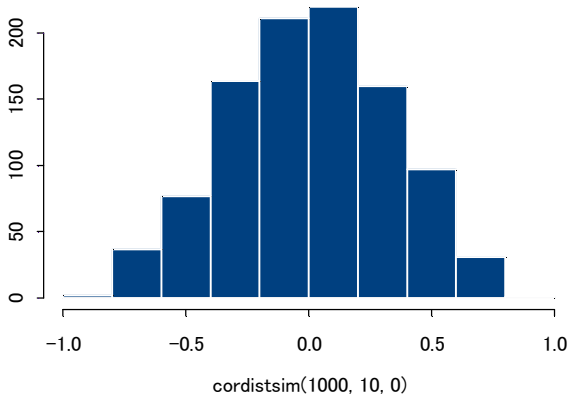
- 母相関係数ρ、標本数N

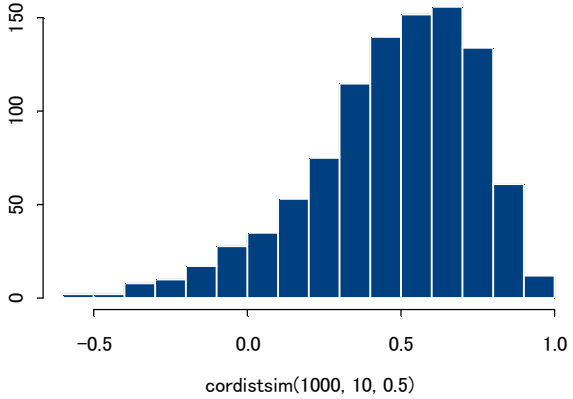
$$f(r)=\frac{2^{n-2}(1-\rho^2)^{n/2}(1-r^2)^{(n-3)/2}}{(n-2)!\pi}\sum_{\alpha=0}^{\infty}\frac{(2\rho r)^{\alpha}}{\alpha!}\Gamma^2[(n+\alpha)/2],\qquad -1\le r\le 1$$

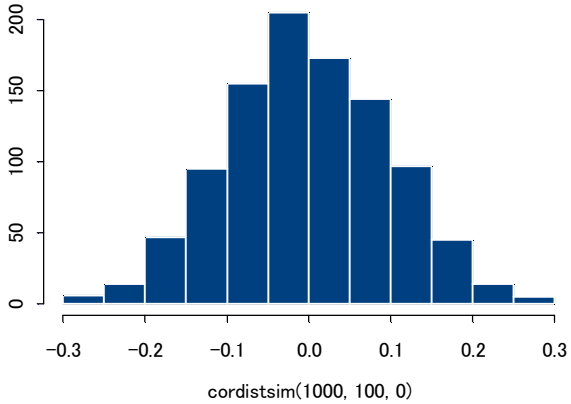
where

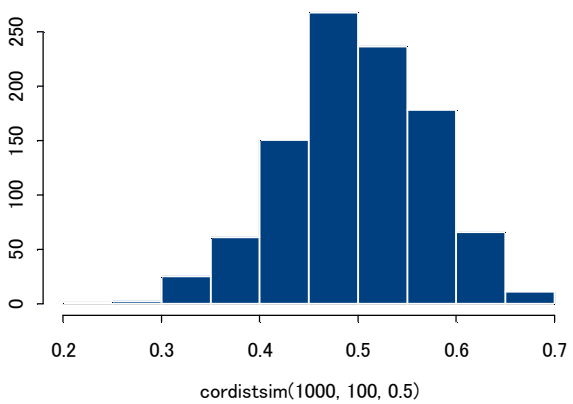
$n=N-1$

```
> cordistsim <- function(nsim, n, r){
+   rslt <- c()
+   for (i in 1:nsim){
+     rdat <- normal2rand(n, r)
+     rslt <-c(rslt, cor(rdat$z1, rdat$z2))
+   }
+   rslt
+ }
> par(mfrow = c(2,2))
> hist(cordistsim(1000, 10, 0))
> hist(cordistsim(1000, 10, 0.5))
> hist(cordistsim(1000, 100, 0))
> hist(cordistsim(1000, 100, 0.5))
```


0
50
100
150
200
−1.0
−0.5
0.0
0.5
1.0
cordistsim(1000, 10, 0)


0
50
100
150
−0.5
0.0
0.5
1.0
cordistsim(1000, 10, 0.5)


0
50
100
150
200
−0.3
−0.2
−0.1
0.0
0.1
0.2
0.3
cordistsim(1000, 100, 0)

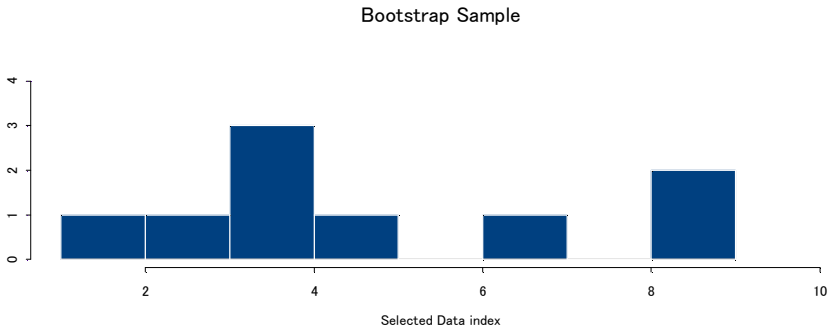

0
50
100
150
200
250
0.2
0.3
0.4
0.5
0.6
0.7
cordistsim(1000, 100, 0.5)

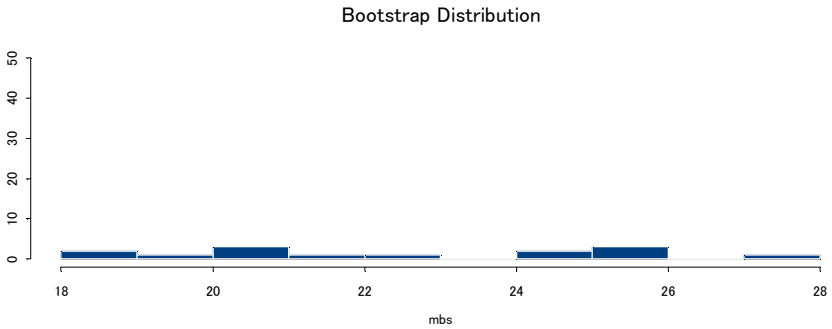
bootstrapデモ

- 関数（R用）、使い方
  - http://case.f7.ems.okayama-u.ac.jp/simulation/bootstrap/bootstrap_r.html
- 関数（S-PLUS用）
  - http://face.f7.ems.okayama-u.ac.jp/~t2/pukiwiki/index.php?bootstrap

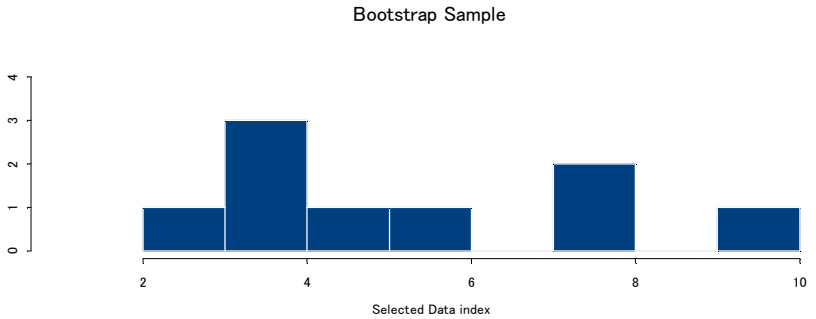
bootstrapdemo

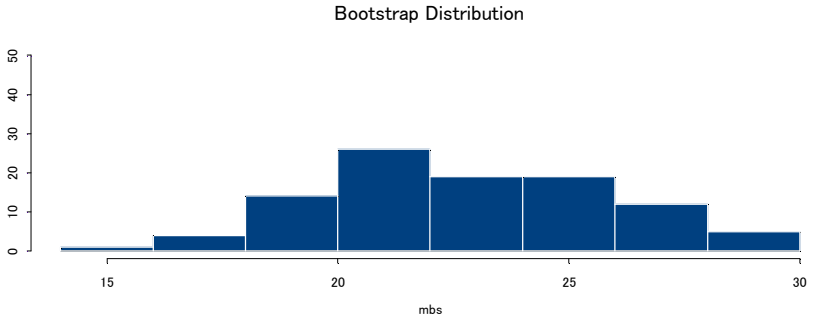
```
> x<-c(1,23,23,32,41,22,18,19,23)
> bootstrapdemo(x, 100)
```



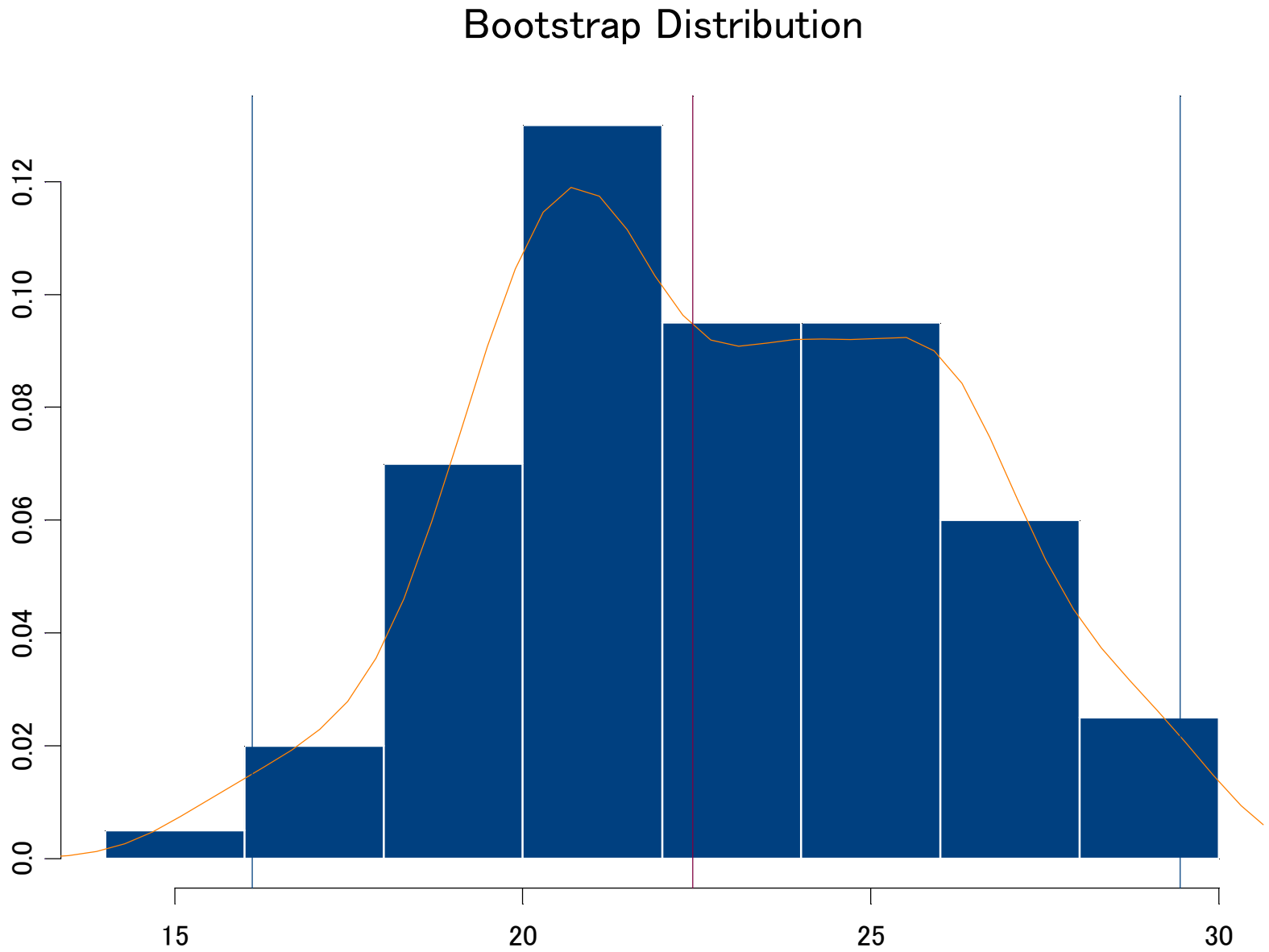


10回目





100回目


Bootstrap Distribution
0.0
0.02
0.04
0.06
0.08
0.10
0.12
15
20
25
30