による統計解析

三中 信宏

minaka@affrc.go.jp
http://cse.niaes.affrc.go.jp/minaka/

〒 305-8604 茨城県つくば市観音台 3-1-3
独立行政法人 農業環境技術研究所 生態系計測研究領域 上席研究員
東京大学大学院 農学生命科学研究科 教授［生態系計測学］
京都大学大学院理学研究科　連携併任教授［進化生物学］
東京農業大学大学院 農学専攻 客員教授［応用昆虫学］

【１】Ｒへの入門

統計言語「Ｒ」は，誰でも利用できる無料のフリーソフトであって，しかも信頼のおけるソフトウェアです．オンラインでのドキュメントがたくさんあり，下記はそのいくつかです．日本語による解説や事例紹介もありますので，参考になると思います．

○ Ｒをインストールする手順については下記を参照してください：

Ｒ - 事始め（Macintosh 版のインストール）
http://aoki2.si.gunma-u.ac.jp/R/begin.html

Ｒ - インストール（Windows/Linux 版のインストール）
http://datamining.tama.ac.jp/~yama/R/install.html

○ Ｒのマニュアルも翻訳されています：

Notes on R (翻訳)
http://datamining.tama.ac.jp/~yama/R/notes.html

○ R - 統計解析とグラフィックスの環境
http://datamining.tama.ac.jp/~yama/R/

○ Rによる統計解析
http://aoki2.si.gunma-u.ac.jp/R/

○ Rの本家である CRAN のウェブサイト：

The Comprehensive R Archive Network
http://cran.r-project.org/

にあるソフトウェア集：

Packaged Softwares
http://cran.r-project.org/src/contrib/PACKAGES.html

にも膨大な数の「R」プログラムがあります.

○ RjpWiki ——Rのための情報交換サイト（※〈R〉の最新情報はここに集結する）
http://www.okada.jp.org/RWiki/index.php?RjpWiki

○ Yahoo! JAPAN ——R関連のヤフー登録サイト一覧
http://dir.yahoo.co.jp/Science/Mathematics/Statistics/Software/R/

○ 租界Rの門前にて——統計言語「R」との極私的格闘記録［三中信宏］
http://cse.niaes.affrc.go.jp/minaka/R/R-top.html

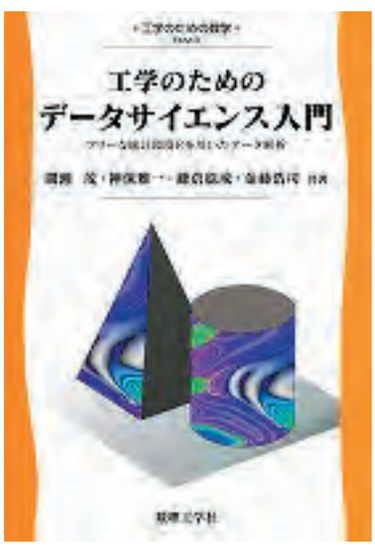
【2】教科書と参考書

昨年から今年にかけて，日本語での〈R〉の解説書が立て続けに出されています．下記の参考書リストはすでに古くなっていますので，最新のリストは私のサイトをごらん下さい：

著者 舟尾暢男
書名 The R Tips：データ解析環境 R の基本技・グラフィックス活用集
刊行 2005 年 3 月 1 日

出版 九天社，東京
頁数 xvi+383 pp. ※ CD-ROM 付
定価 3,500 円（本体価格）
ISBN 4-86167-039-X

著者 岡田昌史
書名 The R Book：データ解析環境 R の活用事例集
刊行 2004 年 6 月 1 日
出版 九天社，東京
URL http://www.okada.jp.org/RWiki/?The%20R%20Book
（RjpWiki の中）
頁数 xii+435 pp. ※ CD-ROM 付
定価 3,800 円（本体価格）
ISBN 4-901676-97-0



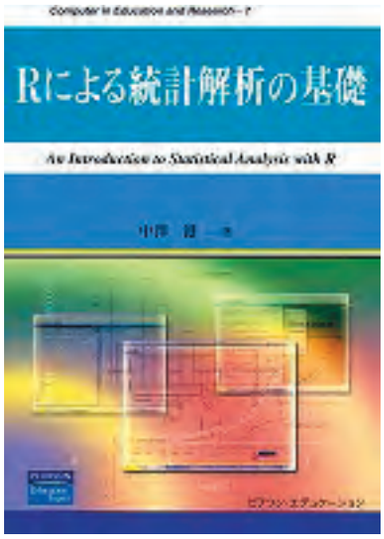
備考「 正誤表 」が公開されています［3/June/2004］.

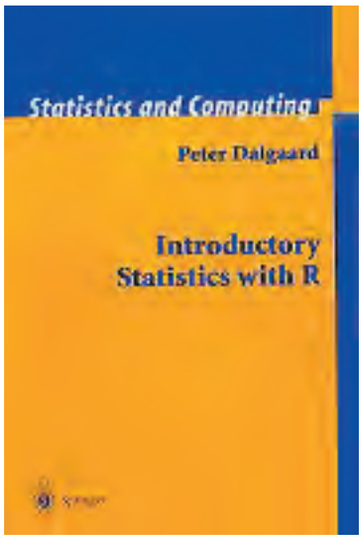
著者 間瀬茂・神保雅一・鎌倉稔成・金藤浩司
書名 工学のためのデータサイエンス入門：フリーな統計環境 R を用いたデータ解析
叢書 工学のための数学 EKM-3
刊行 2004 年 3 月 25 日
出版 数理工学社，東京
頁数 viii+254 pp.
定価 2,300 円（本体価格）
ISBN 4-901683-12-8

編者 東京大学生物測定学研究室
書名 実践生物統計学：分子から生態まで
刊行 2004 年 3 月 1 日
出版 朝倉書店，東京
頁数 viii+136 pp.
定価 3,200 円（本体価格）
ISBN 4-254-42027-7

著者 中澤港
書名 R による統計解析の基礎
叢書 Computer in Education and Research 7
刊行 2003 年 10 月 15 日

出版 ピアソン・エデュケーション，東京
頁数 viii+174 pp.
定価 1,800 円（本体価格）
ISBN 4-89471-757-3
備考「 正誤表 」が公開されています［3/June/2004］.

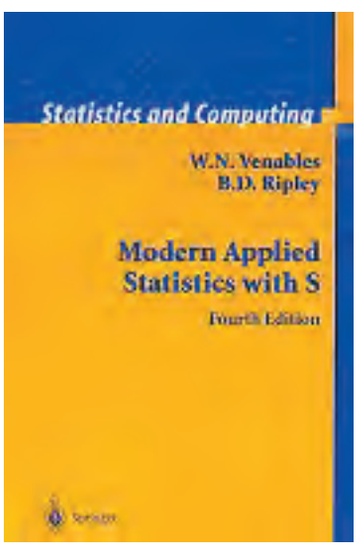




英語の教科書としては 2002 年に出版された下記の本があります：

著者 Peter Dalgaard 2002.
書名 Introductory Statistics with R
刊行 2002 年
出版 Springer Verlag, New York
頁数 xvi+267 pp.
価格 EURO 29.95 (paperback)
ISBN 0-387-95475-9

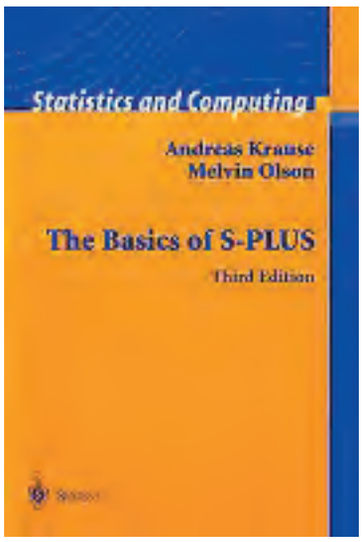
これ↑は、中澤さんの上記の本と併せて，私がテキストとして使っている本です。「R」を横に立ち上げながら使える実習書として利用できます。「R」の機能などひととおりのことはさらっと把握できます。.



Rは，商業ベースのSあるいはS -plus とほぼ同一なので， S（S -plus）の本はそのまま使えます． SやS -plus は日本語の本もすでに出ているので，オンライン検索すればヒットするはず．ここでは，私が持っている新刊を２冊紹介します：

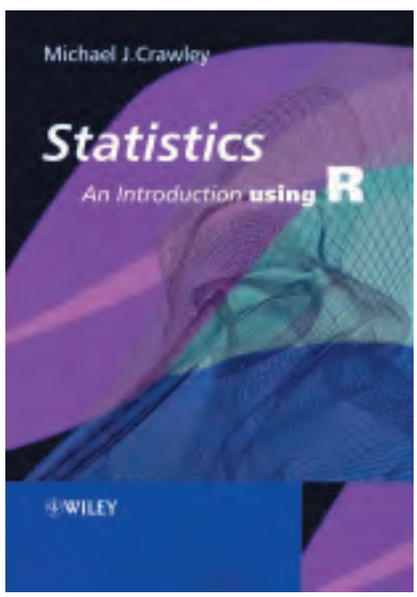
著者 W.N. Venables and B.D. Ripley
書名 Modern Applied Statistics with S (Fourth Edition)
刊行 2002 年
出版 Springer Verlag, New York
頁数 xii+495 pp.
価格 EURO 74.95 (hardcover)
ISBN 0-387-95457-0

この↑本は， Sの定評ある教科書で，第３版はすでに翻訳されています（訳文に難ありとのこと）. タイトルは「S」と銘打たれていますが，中では「S専用」とか「Rでも可能」と付箋されているので， Rのテキストとして利用できます．複雑なデータ解析へのSやRの利用法が書かれているので，実践的に役立ちます.



著者 A. Krause and M. Olsen
書名 The Basics of S-Plus (Third Edition)
刊行 2002 年
出版 Springer Verlag, New York
頁数 xx+419 pp.
価格 EURO 59.95 (paperback)
ISBN 0-387-95456-2

これ↑は S の入門書． R についての章もあります．



なお, つい最近, R による統計学の入門テキストが出ましたので, 最後に付記しておきます．学部生を対象とする教科書です．

著者 Michael J. Crawley
書名 Statistics : An Introduction Using R
刊行 2005 年 4 月
出版 John Wiley & Sons, Chichester
頁数 xiv+327 pp.
価格 US$ 40.00 (paperback)
ISBN 0-470-02298-1
備考 コンパニオンサイト http://www.bio.ic.ac.uk/research/crawley/statistics/

S-Plus の教科書を書いた Crawley による R の新刊教科書です．R 統計学のイントロを学ぶ教材として適しているかも．

【3】データの読みこみ

R にデータを入力するには：

1）コマンドラインからの入力
2）ファイルからの入力

の二つの方法があります．

1）コマンドラインからの入力
簡単なデータならばコマンドラインから直接キー入力してもいいでしょう．たとえば：

```
> daily.intake <- c(5260,5470,5640,6180,6390,6515,6805,7515,7515,8230,8770)
```

※ 11 個の数値データを daily.intake に格納する（「c( )」はベクトル）

```
> daily.intake
[1]  5260 5470 5640 6180 6390 6515 6805 7515 7515 8230 8770
```

※ daily.intake の内容表示

このように入力されたデータについては, たとえば, 平均（mean）・標準偏差（sd）・分位点（quantile）などの記述統計量を下記のように計算できます：

```
> mean(daily.intake)
[1] 6753.636
> sd(daily.intake)
[1] 1142.123
> quantile(daily.intake)
  0%  25%  50%  75% 100%
5260 5910 6515 7515 8770
```

また，母平均 μ に関する仮説を t 検定することも：

```
> t.test(daily.intake, mu=7725)

        One Sample t-test

data:  daily.intake
t = -2.8208, df = 10, p-value = 0.01814
alternative hypothesis: true mean is not equal to 7725
95 percent confidence interval:
 5986.348 7520.925
sample estimates:
mean of x
 6753.636
```

2）ファイルからの読みこみ
しかし，大きなデータの場合には，事前にデータ・ファイルとして別に作成するのが得策です． R に入力できるファイルの形式はプレーンテキストだけです．たとえば，Box1 演習に用いたデータ

ファイル "**Box1_R.tab**" は下記のような内容です：

```
    TRT DATA
001 DM1 2537
002 DM1 2069
003 DM1 2104
004 DM1 1797
005 DM2 3366
006 DM2 2591
007 DM2 2211
008 DM2 2544
009 DDT 2536
010 DDT 2459
011 DDT 2827
012 DDT 2385
013 AZO 2387
014 AZO 2453
015 AZO 1556
016 AZO 2116
017 DB  1997
018 DB  1679
019 DB  1649
020 DB  1859
021 DK  1796
022 DK  1704
023 DK  1904
024 DK  1320
025 Con 1401
026 Con 1516
027 Con 1270
028 Con 1077
```

このデータを Box1 に読みこんで表示させると下記のようになります：

```
> Box1 <- read.table("Box1_R.tab")
> Box1
   TRT DATA
1  DM1 2537
2  DM1 2069
```

```
3   DM1 2104
4   DM1 1797
5   DM2 3366
////////// 中略
28 Con 1077
```

しかし，行番号をあらかじめもたなくても読みこむことができます．たとえば，つぎの "**Box1_R.data**"：

```
TRT DATA
DM1 2104
DM1 1797
DM2 3366
DM2 2591
DM2 2211
DM2 2544
DDT 2536
DDT 2459
DDT 2827
DDT 2385
AZO 2387
AZO 2453
AZO 1556
AZO 2116
DB  1997
DB  1679
DB  1649
DB  1859
DK  1796
DK  1704
DK  1904
DK  1320
Con 1401
Con 1516
Con 1270
Con 1077
```

をＲに入力するには，下記の指定をします：

> Box1 <- read.table("Box1_R.data", header=T)

「header=T」とは，第 1 行に列の名称（ヘッダ）が指定されているという意味で，この読みこみをした結果は，上の場合と同一です：

```
   TRT DATA
1  DM1 2537
2  DM1 2069
3  DM1 2104
///////////
28 Con 1077
```

ほかに，Excel のデータシート（**Box1_R.xls**）から読み込む場合，いったん Excel で「csv」形式（**Box1_R.csv**）で保存しなおした上で，

> Box1 <- read.csv("Box1_R.csv", header=T)

とすれば，そのデータを R に読み込むことができます．また，データがタブ区切りされている場合は「**read.delim**」コマンドで同様に読み込めます．

なお，ひとまとまりの計算が終了するたびに，作業スペースを掃除しておくと，思わぬエラーを回避できます．

> detach()　　#「**detach**」したものをすべてはずす．
> rm(list=ls())　# すべてのオブジェクトを消去する．

【 4 】正規分布に関連する関数（dnorm, pnorm, qnorm, rnorm）

●平均 0, 標準偏差 0.8 の正規分布の確率密度関数（dnorm）

> x <- seq(-3, 3, 0.05)
> plot(x,dnorm(x, mean=0, sd=0.4), type="n")
> curve(dnorm(x, mean=0, sd=0.8), type="l",add=T)

●正規分布の確率分布関数（pnorm）とその逆関数（qnorm）

> curve(pnorm(x, mean=0, sd=0.8), type="l", lty=3, add=T)

● 5 %点の表示

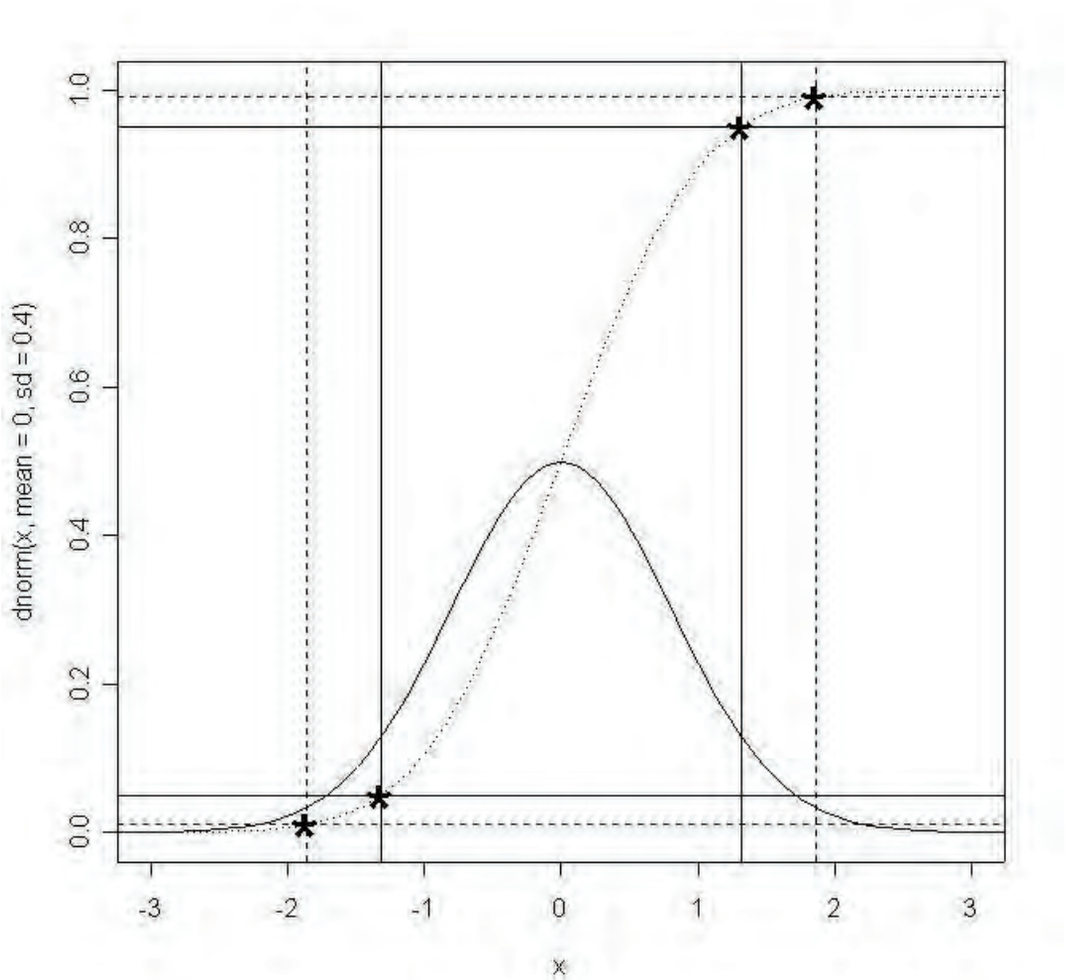
> abline(h=0.05)

```
> lower.alpha5 <- qnorm(0.05, mean=0, sd=0.8)
> lower.alpha5
 [1] -1.315883
> abline(v=lower.alpha5)
> points(lower.alpha5, 0.05, cex=3.0, pch="*")

> abline(h=0.95)
> upper.alpha5 <- qnorm(0.05, mean=0, sd=0.8, lower.tail = FALSE)
> upper.alpha5
 [1] 1.315883
> abline(v=upper.alpha5)
> points(upper.alpha5, 0.95, cex=3.0, pch="*")
```

● 1 %点の表示

```
> abline(h=0.01, lty=2)
```

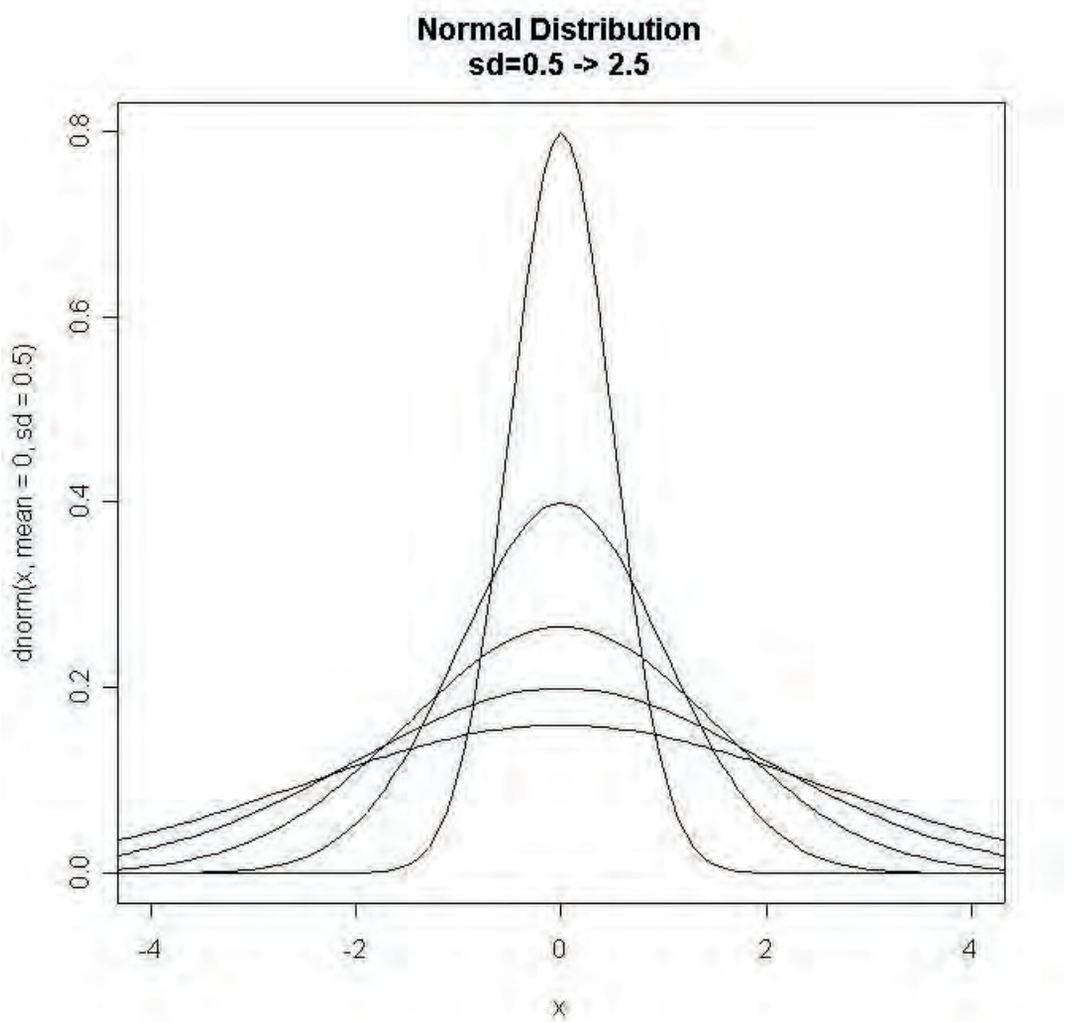


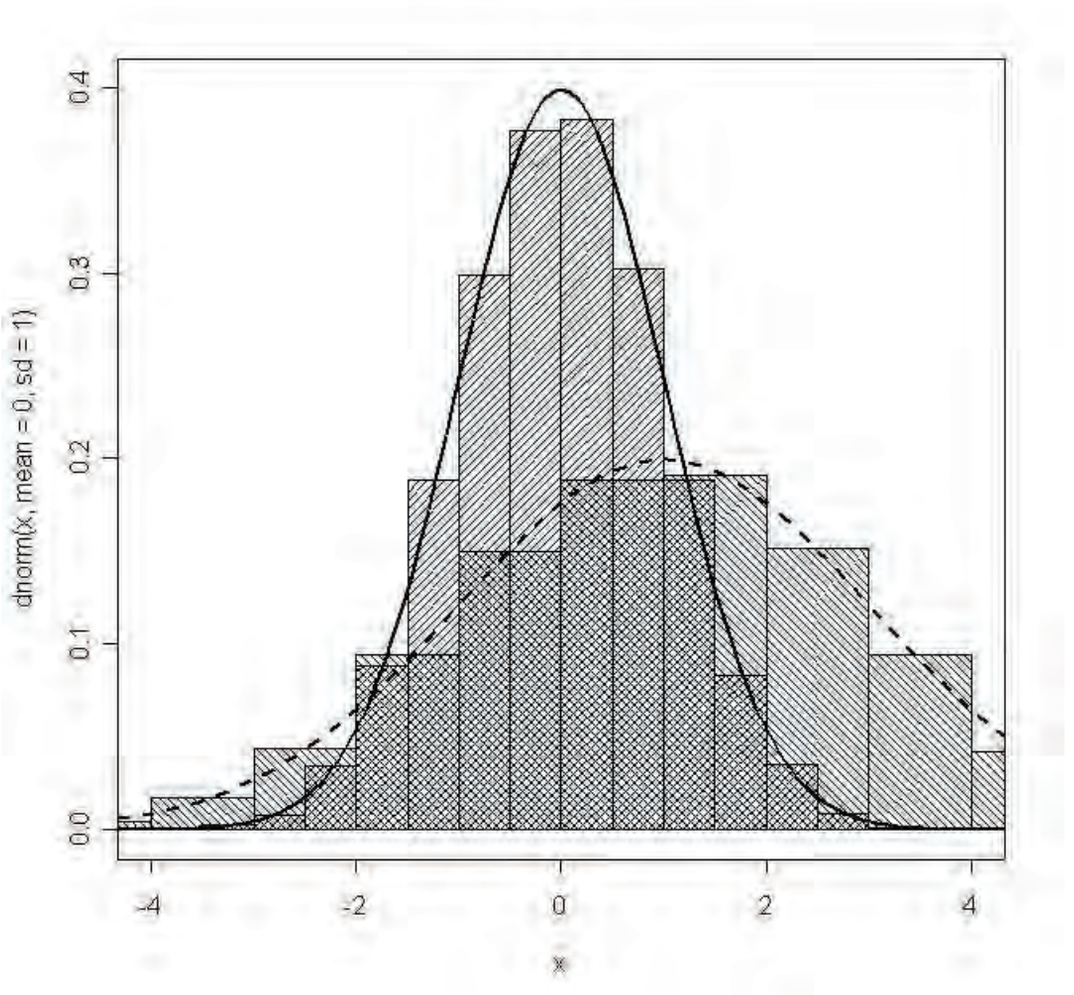
```
> lower.alpha1 <- qnorm(0.01, mean=0, sd=0.8)
> lower.alpha1
 [1] -1.861078
> abline(v=lower.alpha1, lty=2)
> points(lower.alpha1, 0.01, cex=3.0, pch="*")

> abline(h=0.99, lty=2)
> upper.alpha1 <- qnorm(0.01, mean=0, sd=0.8, lower.tail = FALSE)
> upper.alpha1
 [1] 1.861078
> abline(v=upper.alpha1, lty=2)
> points(upper.alpha1, 0.99, cex=3.0, pch="*")
```

●正規乱数（rnorm）の生成とヒストグラム表示

```
> random.norm <- rnorm(10, mean=0, sd=0.8)
> hist(random.norm, freq=F)
```






```
> random.norm <- rnorm(100, mean=0, sd=0.8)
> hist(random.norm, freq=F)
> random.norm <- rnorm(1000, mean=0, sd=0.8)
> hist(random.norm, freq=F)
> random.norm <- rnorm(10000, mean=0, sd=0.8)
> hist(random.norm, freq=F)
> random.norm <- rnorm(100000, mean=0, sd=0.8)
> hist(random.norm, freq=F)
> random.norm <- rnorm(1000000, mean=0, sd=0.8)
> hist(random.norm, freq=F)
> curve(dnorm(x, mean=0, sd=0.8), add=T)
```

●正規分布のパラメーター（１）――平均μを変える

```
> x <- seq(-4, 4, 0.01)
> plot(x, dnorm(x, mean=0, sd=0.5), type="n")
> title("Normal Distribution¥nmean=0 -> 2.0")
```

> for (i in 1:5) curve(dnorm(x, mean=0.5*(i-1), sd=0.5), type="l", add=T)

●正規分布のパラメーター（２）――分散σ^2 を変える

> x <- seq(-4, 4, 0.01)

> plot(x, dnorm(x, mean=0, sd=0.5), type="n")

> title("Normal Distribution¥nsd=0.5 -> 2.5")

> for (i in 1:5) curve(dnorm(x, mean=0, sd=0.5*i), type="l", add=T)

●標準正規分布（平均０，分散１）

> mean1 <- 1.0

> sd2 <- 2.0

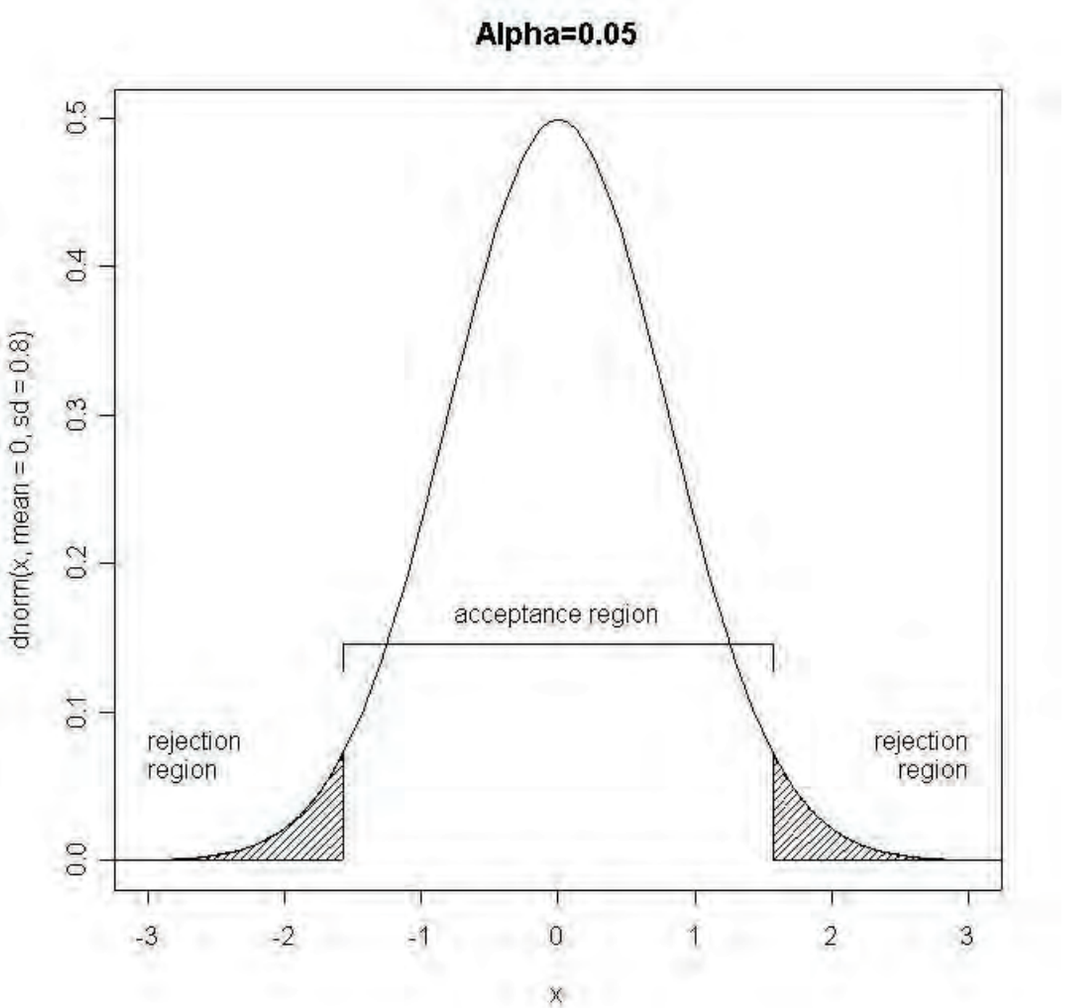
> plot(x, dnorm(x, mean=0, sd=1), type="n")

> x <- rnorm(10000, mean=mean1, sd=sd2)

> hist(x, freq=F, density=25, angle=135, add=T)

> curve(dnorm(x, mean=mean1, sd=sd2), type="l", lty=2, lwd=2, add=T)

> hist((x - mean1)/sd2, freq=F, density=25, angle=45, add=T)



```
> curve(dnorm(x, mean=0, sd=1), type="l", lty=1, lwd=2, add=T)
```

【５】正規分布のもとでの棄却域の図示

●正規分布（平均 0, 標準偏差 0.8）の図示

```
> x <- seq(-3,3,0.01)
> plot(x,dnorm(x,mean=0,sd=0.8),type="n")
> curve(dnorm(x,mean=0,sd=0.8),type="l",add=T)
```

●棄却水準（$\alpha = 0.05$）を設定と表示

```
> alpha <- 0.05
> title("Alpha=0.05")
```

●左側棄却域の表示

```
> xmin <- -3
> xmax <- 3
> critical.left <- qnorm(alpha/2, mean=0, sd=0.8)
> xaxis <- seq(xmin, critical.left, length=100)
> yaxis <- c(dnorm(xaxis, mean=0, sd=0.8), 0, 0)
> xaxis <- c(xaxis, critical.left, xmin)
> polygon(xaxis, yaxis, density=25)
```

●右側棄却域の表示

```
> critical.right <- qnorm(alpha/2, mean=0,sd=0.8,lower.tail=F)
> xaxis <- seq(critical.right, xmax, length=100)
> yaxis <- c(dnorm(xaxis, mean=0, sd=0.8), 0, 0)
> xaxis <- c(xaxis, xmax, critical.right)
> polygon(xaxis, yaxis, density=25)
```

●棄却域タイトル表示

```
> ypos <- dnorm(critical.left, mean=0, sd=0.8)
> text(xmin, ypos, "rejection¥nregion", adj=0)
> text(xmax, ypos, "rejection¥nregion", adj=1)
```

●受容域タイトル表示

```
> text((critical.left+critical.right)/2, 2*ypos+0.02, "acceptance region")
> xaxis <- c(rep(critical.left,2), rep(critical.right,2))
> yaxis <- c(2*ypos-0.02, 2*ypos, 2*ypos, 2*ypos-0.02)
> lines(xaxis,yaxis)
```

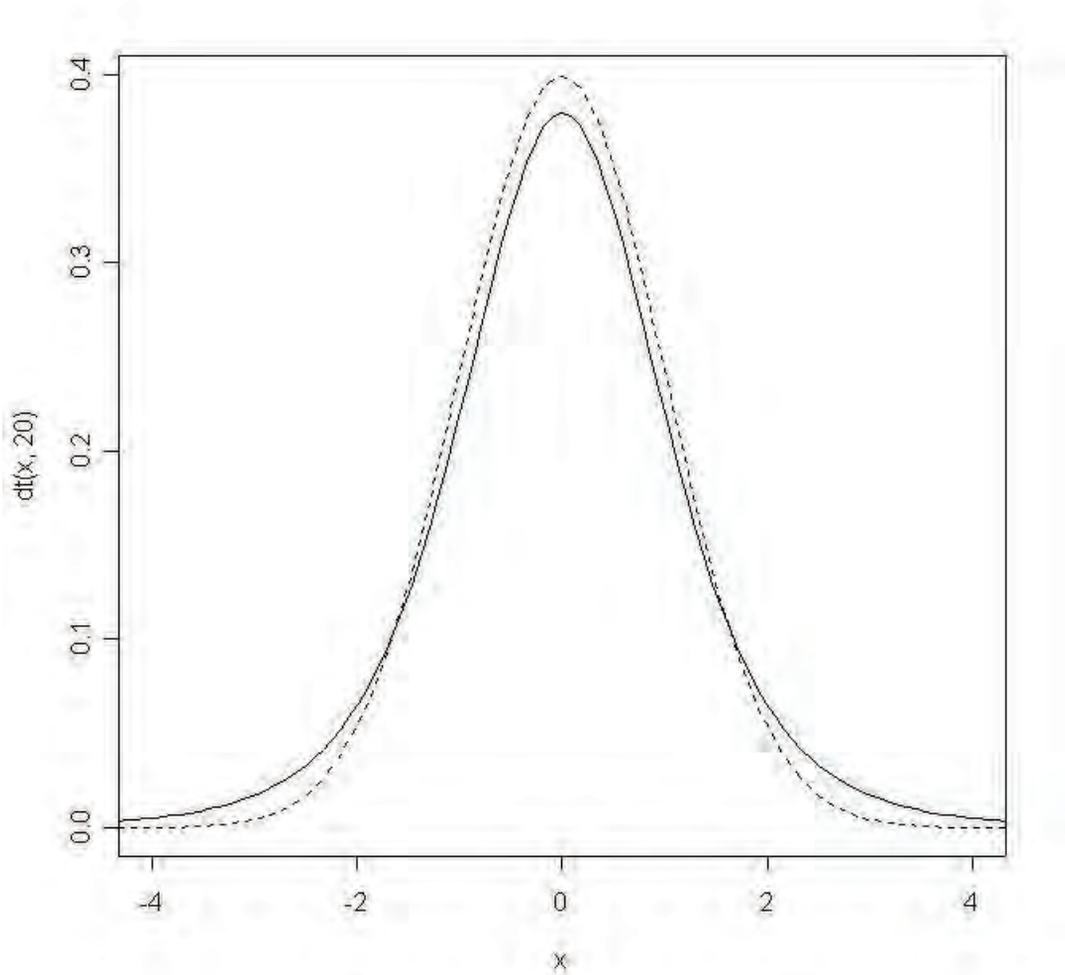
● $\alpha = 0.05$ での棄却水準値

```
> critical.left
[1] -1.567971
> critical.right
[1] 1.567971
```

【6】 t 分布に関連する関数（dt, pt, qt, rt）

● t 分布の密度関数（dt）を表示死，標準正規分布と比較

```
> x <- seq(-4, 4, 0.01)
> plot(x, dt(x, 20), type="n")
> curve(dt(x, 5), type="l", add=T)
> curve(dnorm(x), type="l", lty=2, add=T)
```

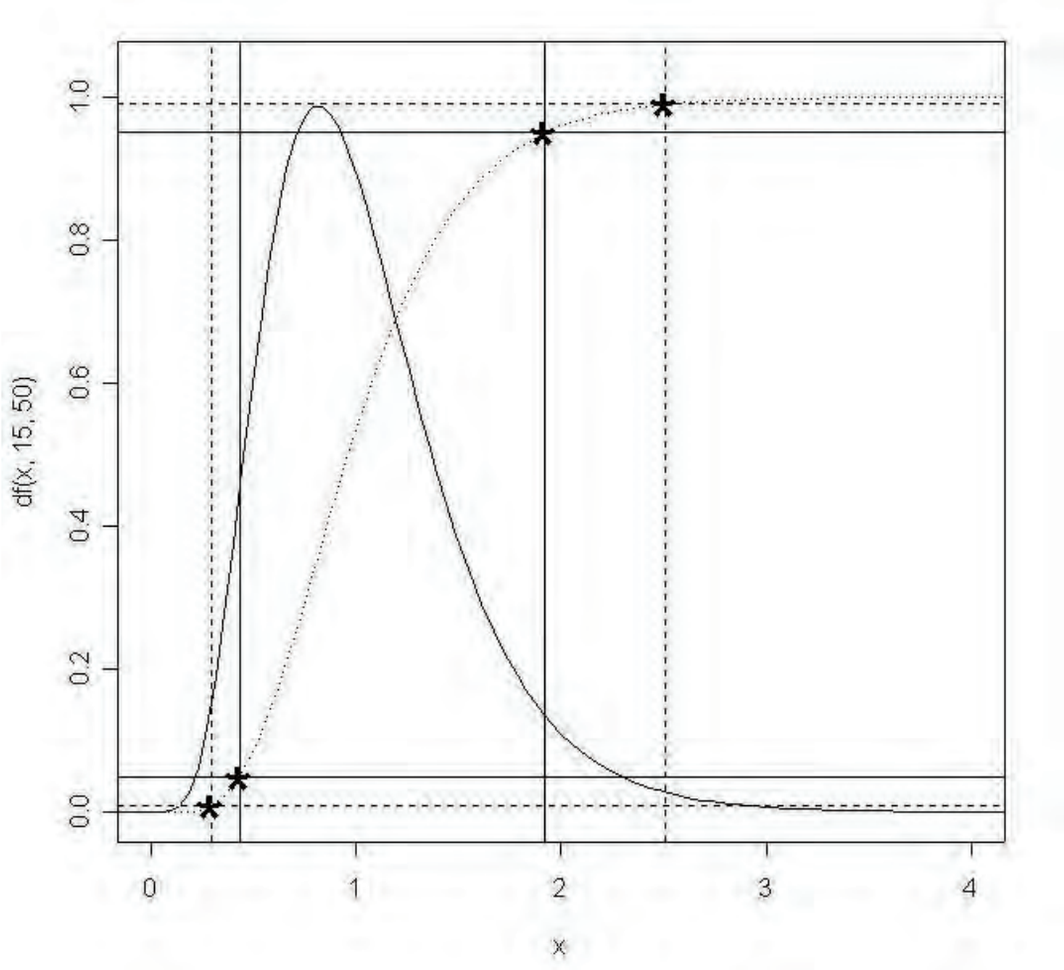


● 5 %点の表示

```
> abline(h=0.05)
> lower.alpha5 <- qt(0.05, 5)
> lower.alpha5
 [1] -2.015048
> abline(v=lower.alpha5)
> points(lower.alpha5, 0.05, cex=3.0, pch="*")
> upper.alpha5 <- -lower.alpha5
> upper.alpha5
 [1] 2.015048
> abline(v=upper.alpha5)
> points(upper.alpha5, 0.95, cex=3.0, pch="*")
```

● t 分布のパラメーター――自由度を変える

```
> x <- seq(-4, 4, 0.01)
```



```
> plot(x, dt(x, 20), type="n")
> title("t Distribution¥ndf=5 -> 1")
> for (i in 1:5) curve(dt(x, 5-(i-1)), type="l", add=T)
```

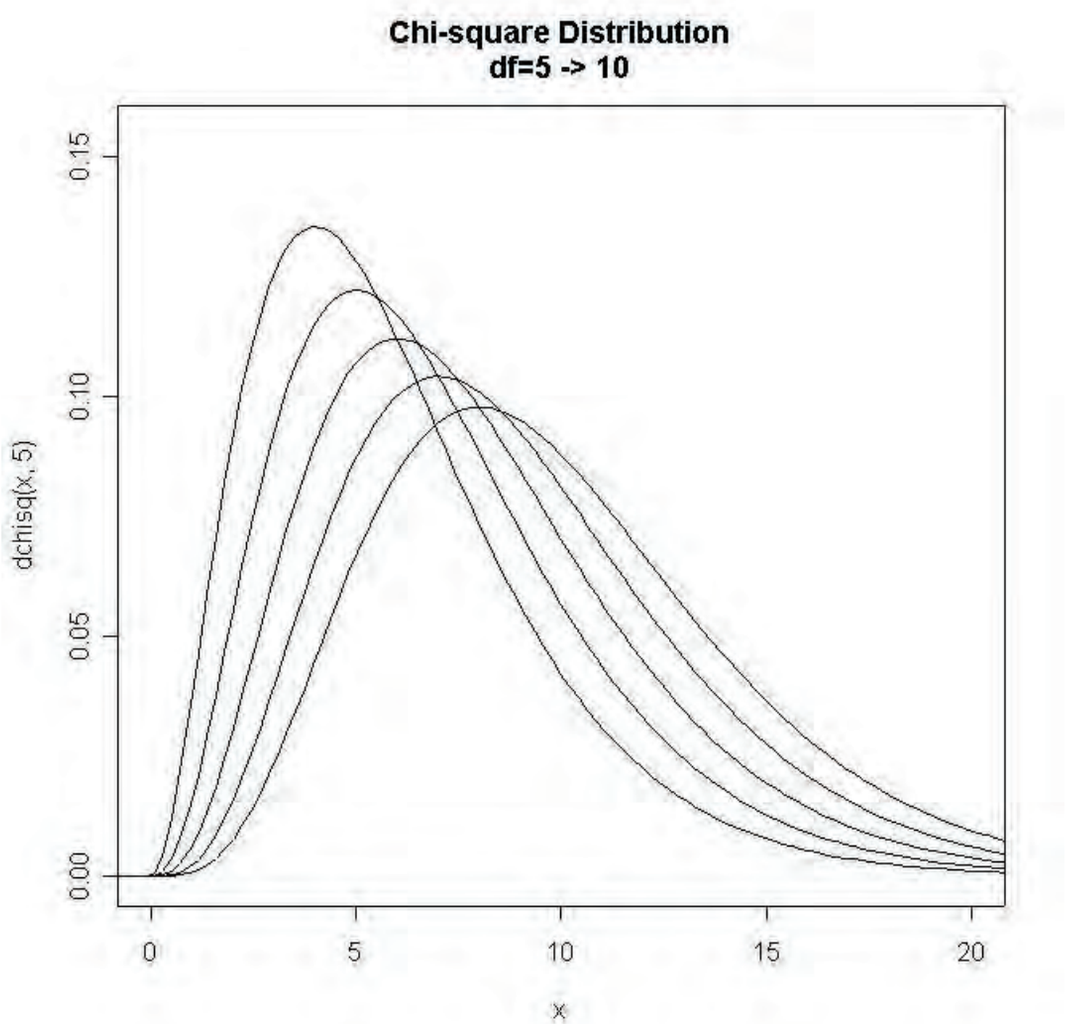
【7】χ二乗分布に関連する関数（dchisq, pchisq, qchisq, rchisq）

● χ二乗分布の密度関数（dchisq）を表示

```
> x <- seq(0, 20, 0.01)
> plot(x, dchisq(x, 5), type="n")
> curve(dchisq(x, 10), type="l", add=T)
```

● t 分布のパラメーター――自由度を変える

```
> x <- seq(0, 20, 0.01)
```



```
> plot(x, dchisq(x, 5), type="n")
> title("Chi-square Distribution¥ndf=5 -> 10")
> for (i in 1:5) curve(dchisq(x, 5+i), type="l", add=T)
```

【８】F 分布に関連する関数（df, pf, qf, rf）

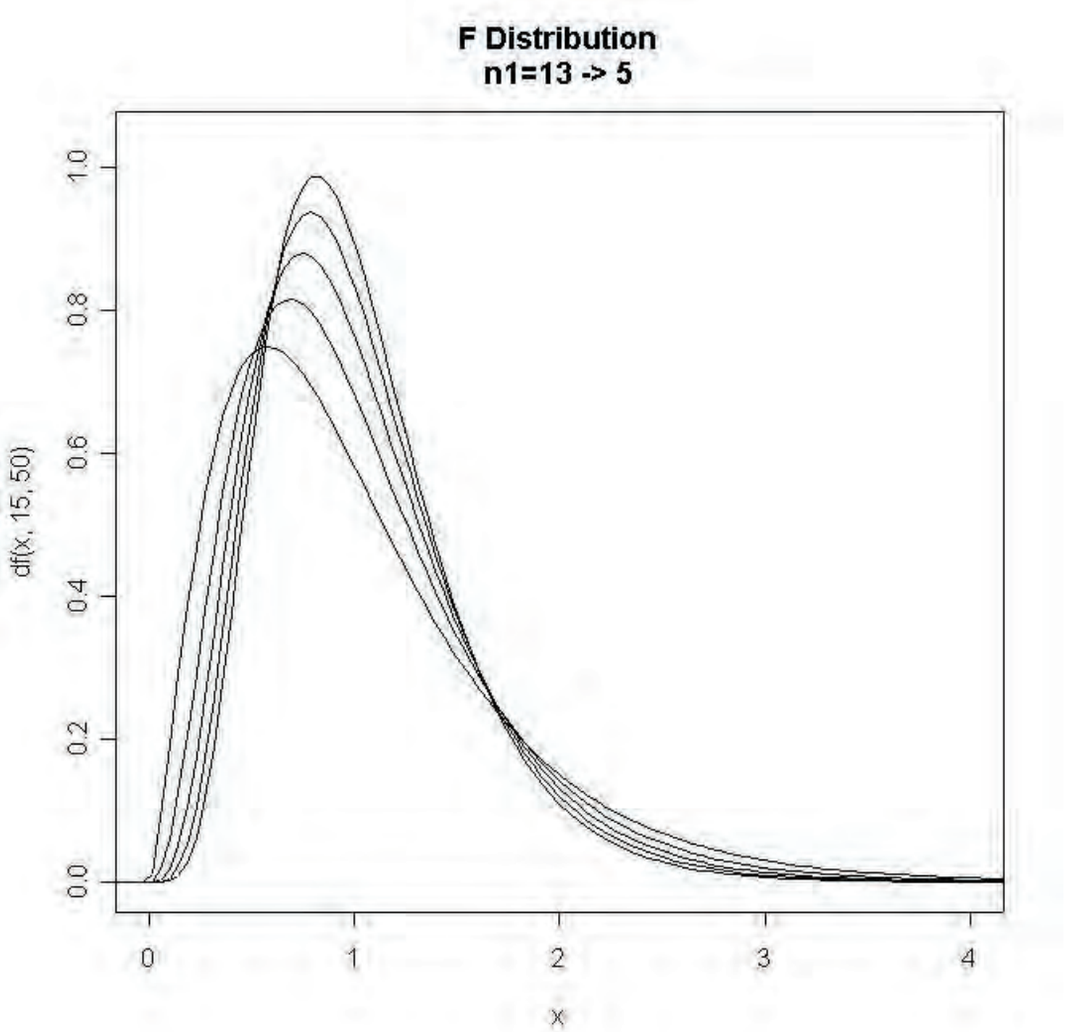
● F 分布の密度関数（df）を表示

```
> x <- seq(0, 4, 0.01)
> plot(x, df(x, 15, 50), type="n")
> curve(df(x, 13, 50), type="l", add=T)
```

● F 分布の確率分布関数（pf）を表示

```
> curve(pf(x, 13, 50), type="l", lty=3, add=T)
```

● 5 %点の表示

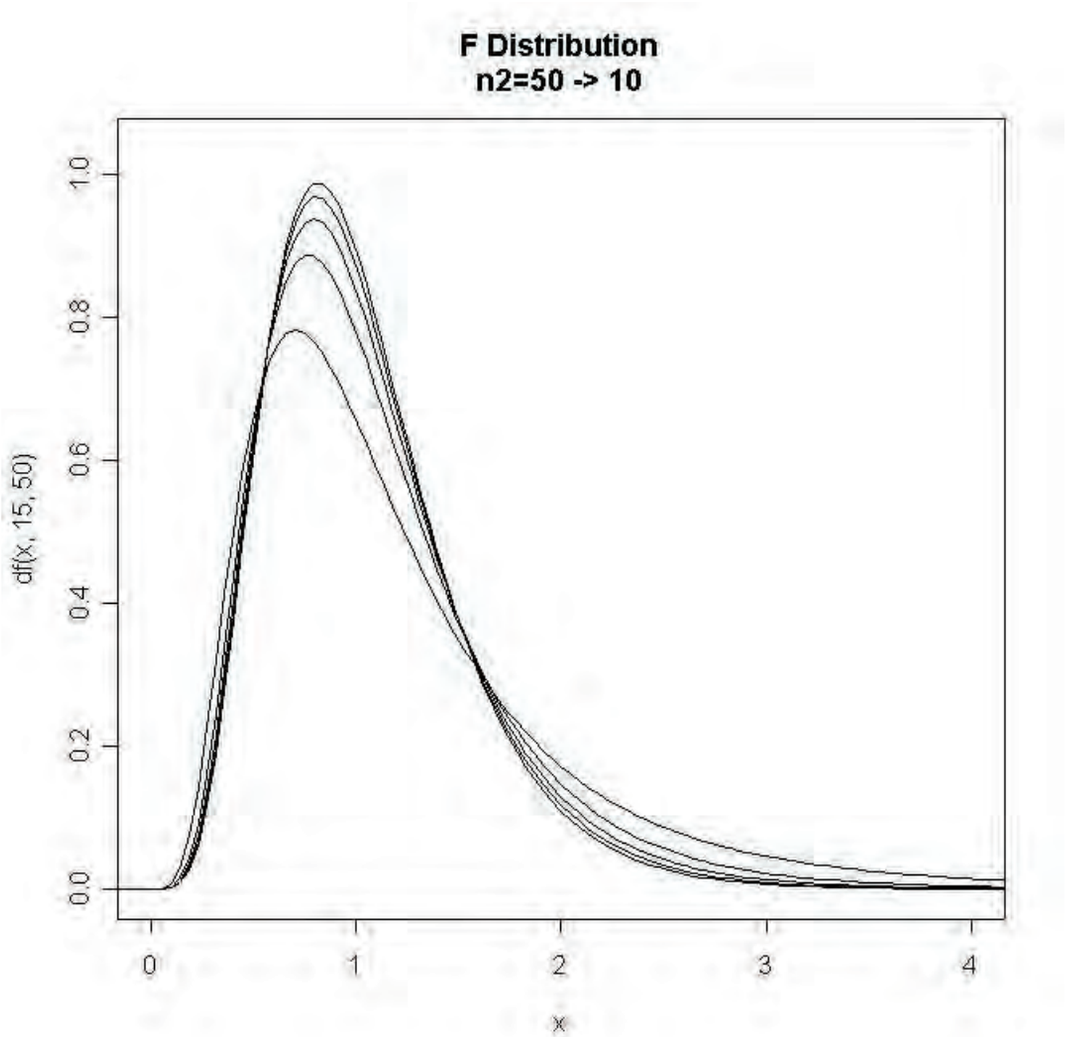


```
> abline(h=0.05)
> lower.alpha5 <- qf(0.05, 13, 50)
> lower.alpha5
 [1] 0.4321874
> abline(v=lower.alpha5)
> points(lower.alpha5, 0.05, cex=3.0, pch="*")

> abline(h=0.95)
> upper.alpha5 <- qf(0.05, 13, 50, lower.tail = FALSE)
> upper.alpha5
 [1] 1.921429
> abline(v=upper.alpha5)
> points(upper.alpha5, 0.95, cex=3.0, pch="*")
```

● 1 %点の表示

```
> abline(h=0.01, lty=2)
```



```
> lower.alpha1 <- qf(0.01, 13, 50)
> lower.alpha1
 [1] 0.2962809
> abline(v=lower.alpha1, lty=2)
> points(lower.alpha1, 0.01, cex=3.0, pch="*")

> abline(h=0.99, lty=2)
> upper.alpha1 <- qf(0.01, 13, 50, lower.tail = FALSE)
> upper.alpha1
 [1] 2.508328
> abline(v=upper.alpha1, lty=2)
> points(upper.alpha1, 0.99, cex=3.0, pch="*")
```

●F 分布のパラメーター（1）――分子自由度 n1 を変える

```
> x <- seq(0, 4, 0.01)
> plot(x, df(x, 15, 50), type="n")
> title("F Distribution¥nn1=13 -> 5")
> for (i in 1:5) curve(df(x, 13-2*(i-1), 50), type="l", add=T)
```

●F 分布のパラメーター（2）――分母自由度 n2 を変える

```
> x <- seq(0, 4, 0.01)
> plot(x, df(x, 15, 50), type="n")
> title("F Distribution¥nn2=50 -> 10")
> for (i in 1:5) curve(df(x, 13, 50-10*(i-1)), type="l", add=T)
```

●F 分布のパラメーター（3）――分子と分母の自由度を同時に変える

```
> x <- seq(0, 4, 0.01)
> plot(x, df(x, 15, 50), type="n")
> title("F Distribution¥nn1=13 -> 5¥nn2=50 -> 10")
> for (i in 1:5) curve(df(x, 13-2*(i-1), 50-10*(i-1)), type="l", add=T)
```