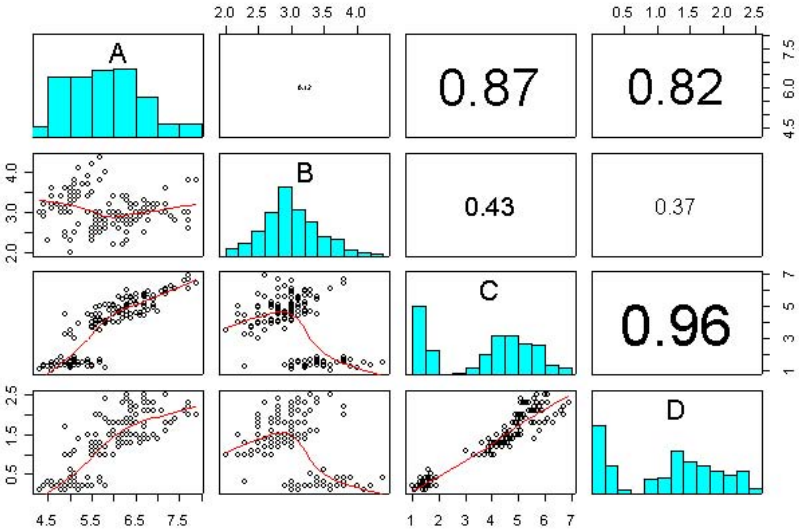
Rによるデータ解析入門

熊澤　吉起



目次

目次 ii

第 1 章　データの種類と図式化 1
- 1.1　データの種類 1
- 1.2　ソフトウェア R のデータ例 1
- 1.3　データの図式化 6
- 問題 13

第 2 章　データの入力と計算 15
- 2.1　ソフトウェア R の起動と終了 15
- 2.2　データの入力 15
- 2.3　データの計算 21
- 2.4　オブジェクトの一覧と削除 21

第 3 章　データの数値化 23
- 3.1　位置の尺度 23
- 3.2　散らばりの尺度 25
- 3.3　データ間の関連 28
- 3.4　データの特性をみる量 28
- 問題 31

第 4 章　一標本正規母集団に関する検定 32
- 4.1　平均の検定 32
- 4.2　母分散 σ^2 の検定 37
- 4.3　対になった標本に関する検定 42
- 問題 43

第 5 章　二標本正規母集団に関する検定 44
- 5.1　平均の差の検定 45
- 5.2　分散の比 σ_1^2/σ_2^2 の検定 50
- 問題 53

第 6 章　適合度の検定 54
- 6.1　度数の適合性 54
- 6.2　分割表と独立性の検定 55
- 6.3　正規性の検定 58
- 6.4　Kolmogorov-Smirnov 検定 59
- 問題 62

第 7 章　相関分析と単回帰分析 63
- 7.1　散布図 63
- 7.2　相関係数の検定 65

7.3 単回帰分析 68
問題 71

第 8 章 品質管理 72
8.1 特性要因図 72
8.2 管理図 72
8.3 チェックシート 79
8.4 層別 79

付録 A 検索 80
A.1 インターネット検索 80
A.2 HDD 内の検索 83

付録 B 仮説検定について 86

付録 C R のインストール 91
C.1 Mac OS X 版 91
C.2 Windows 版 97

付録 D エディタについて 105
D.1 エディタ 105

参考文献 109

ギリシャ文字 110

索引 111
Tips 索引 111

R の関数・データ索引 112

一般・統計用語索引 114

英語索引 116

図目次

1.1 データ sleep 2
1.2 一様乱数のヒストグラム 4
1.3 データ co2 の時系列解析 5
1.4 データ VADeaths 7
1.5 正規乱数のヒストグラム 8
1.6 パレート図 9
1.7 正規確率プロット 9
1.8 100 個の正規乱数の箱ひげ図 10
1.9 アヤメの種ごとの箱ひげ図 11
1.10 アヤメの散布図 11
1.11 データ cars の散布図と回帰直線 12
1.12 データ Titanic のモザイクプロット 12

2.1 時系列解析 20

3.1 歪度 29
3.2 尖度 30

4.1 アヤメの種 setosa のガク片の長さのヒストグラム 33

5.1 データ ToothGrowth のヒストグラムと箱ひげ図 45

6.1 箱ひげ図と正規確率プロット 58
6.2 setosa と versicolor の ecdf 62

7.1 散布図 63
7.2 収入と支出の散布図と回帰直線 69
7.3 cars の speed と dist の散布図と回帰直線 70

8.1 管理図の一例 73

第 1 章　データの種類と図式化

1.1　データの種類

データ(data) は**標本**(sample)、資料とも呼ばれ、個体に対する観測や繰り返し実験を行うことで収集されるものである。最近ではコンビニエンスストアで買い物をしたとき、レジでは品名、単価、数量の他に性別・年齢が入力されています。これに当日の曜日、天候、該当店舗の広告等のデータが追加され、商品の売れ残りを減少させるための発注量の決定に利用されている。

1.1.1　質的データ

質的データ(qualitative data) は、定性的データとも呼ばれ、個体の属性や内容を表すもので文字や記号を用いて記述される。

- **名義尺度**(nominal scale)：分類ために名称を用いて表されるデータ。性別、職業、血液型など。
- **順序尺度**(ordinal scale)：適当な基準・標準により順序づけられた一連のデータ。成績順位、CD 売り上げ順位など。アンケート調査の「好き」「どちらでもない」「嫌い」等も順序尺度となる。

1.1.2　量的データ

量的データ(quantitative data) は、定量的データとも呼ばれ、測定・観測により大きさ、長さ、重さ等の単位を伴い記述されるものであるものである。連続な値をとる**連続型データ**(continuous data) と離散な値をとる**離散データ**(discrete data) に分けられる。身長や体重などのデータは離散値として通常扱われるが、統計学では連続データとして考える。

- **間隔尺度**(interval scale)：数値の間隔に意味があり、原点が指定されていない尺度。偏差値、体温など。
- **比例尺度**(ratio scale)：尺度の原点が定まっており、四則演算が出来るデータ。体重、身長など。

1.2　ソフトウェア R のデータ例

利用しているパソコンにソフトウェア R がインストールされており、これにデフォルトとして含まれているデータについて考える。Mac OS X と Windows でのインストールについては付録 C (91 ページ) 参照。

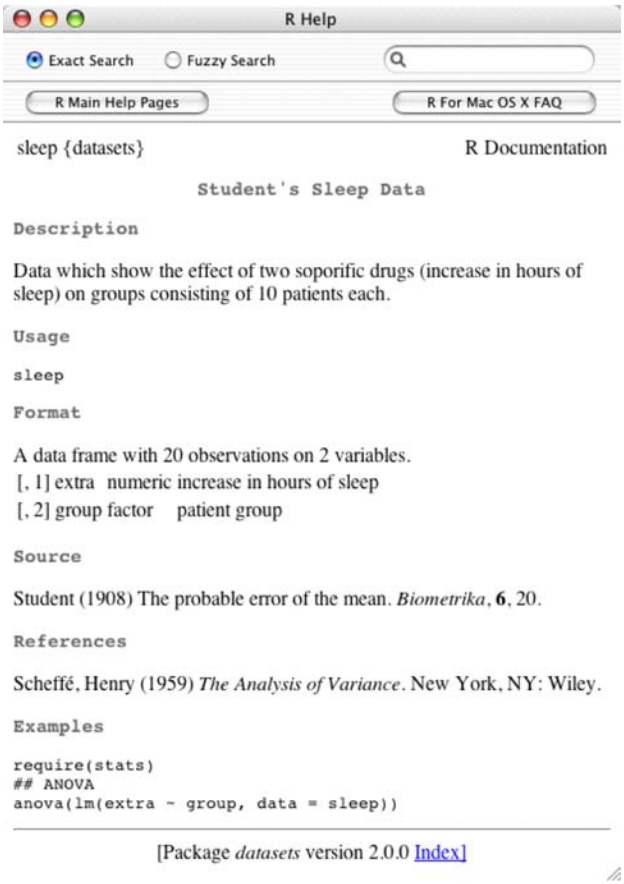
例題 1.1. R のデータ `sleep` を考える。■

入力

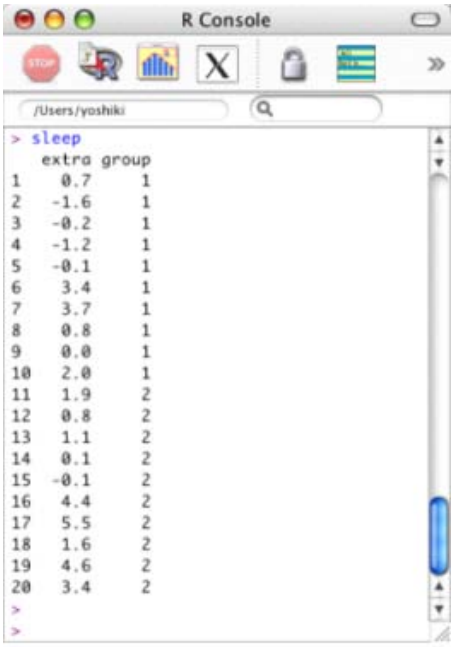
```
> data(sleep)    # データ sleep を組み込む。
> help(sleep)    # データ sleep の説明表示。図 1.1(a) 参照。
> sleep          # データ sleep の内容表示。図 1.1(b) 参照。
```

左の R の関数を入力する際、# 以降は注釈を表しており、入力の際必ずしも必要でない。

Mac OS X での R の出力である図 1.1(a) の `help(sleep)` から分かるように、20 行 (row)、2 列 (column) の大きさで、2 つの群に対する睡眠薬の効果を示している。Windows 版の R でも同じ出力結果が得られます。図 1.1(b) から列名は `extra` と `group` から成り、`extra` は睡眠薬による睡眠時間の増加時間を表し、`group` は患者群を表している。`extra` は時間で比例尺度、`group` は名義尺度となっている。



(a) sleep の説明



(b) sleep の表示

図 1.1　データ sleep

R で用いられる変数 (オブジェクト) 名や関数名は大文字と小文字を区別します。

`Error: Object "foo" not found` などのエラーが出たときはキーボードの上向き矢印キー↑を押すことで、直前までに入力したコマンドが表示され、左向きの矢印キー←を押し修正したい箇所にポインタを移動して、delete キーまたは BS キーを押して不要なものを削除し、正しいコマンドになるように修正できたら、Enter キー↵を押します。また、付録 D にあるようなエディタ起動させておき、エディタにコマンドやデータセットをエディタに入力して、必要な箇所を copy し、R のコンソールに paste してコマンドを実行させることも有効な方法です。

Tips help について — 1

R の関数には同じ役割を果たすものがあり、help と?は同じ結果が得られます。

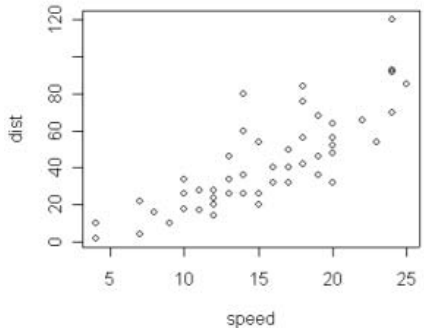
入力

```
> help(sleep)    # データ sleep の説明表示。
> ?sleep         # データ sleep の説明表示。
```

例題 1.2. データ cars を考える。これも R の基本データとして組み込まれており、1920 年代の車の速度 speed と停止に要する距離 dist からなっている。■

入力と出力結果

```
> data(cars)                    # データ cars の組み込み。
> help(cars)                    # データ cars の説明表示。
> cars                          # データ cars の内容表示。
   speed dist
1      4    2
2      4   10
   ⋮    ⋮                          <<途中略>>
49    24  120
50    25   85
> plot(dist~speed, data=cars)   # 右図の speed と dist の散布図。
```



dist speed は dist を speed で表そうという単回帰式[*1]を表しており、plot(dist speed, data=cars) で散布図が描かれる。

例題 1.3. R のデータ Titanic を考える。R の基本データとして組み込まれており、船 Titanic が沈没したときの状態を class (1st、2nd、3rd、Crew)、Sex (Male、Female)、Age (Child、Adult)、Survived (yes、no) の項目で調査してもので、表の形で与えられており、分割表 (contingency table) と呼ばれる[*2]。■

入力と出力結果

```
> data(Titanic)                 # データ Titanic の組み込み。
> ?Titanic                      # データ Titanic の説明表示。
> Titanic                       # データ Titanic の内容表示。
, , Age = Child, Survived = No
      Sex
Class  Male Female
  1st     0     0
  2nd     0     0
  3rd    35    17
  Crew    0     0
              ⋮                 <<途中略>>
  Crew  192    20
```

例題 1.4. 一様乱数について考える。これは、R の関数 runif()[*3]によって生成されるもので、区間 $[0, 1]$ 上の一様分布からの乱数である。■

入力と出力結果

```
> ?runif          # 関数 runif の説明表示。
> x<- runif(100)  # 一様乱数を 100 個発生させ、変数 x に代入。
> hist(x)         # x のヒストグラムの作成。図 1.2 参照。
```

コマンド hist(x) によりオブジェクト x のヒストグラム (histogram)[*4]が作図されます。

[*1] 68 ページ参照。
[*2] 55 ページ参照。
[*3] 乱数の初期設定は set.seed を用います。
[*4] 7 ページ参照。

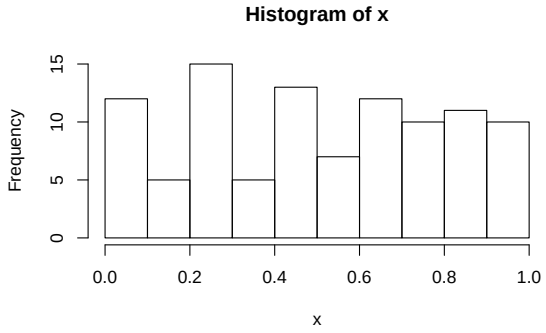


図 1.2　一様乱数のヒストグラム

Tips　乱数について

乱数とは、ランダムな数値の系列のことで、確率論・統計学やゲームの世界では非常に重要な役割を果たすものある。疑似乱数とも呼ばれ、計算機においては、平方採中法、乗算型合同法、相加型合同法、混合型合同法等の手法に基づき発生させます。長い周期を持つ乱数が良い乱数と考えられています。

R にはパッケージと呼ばれるものが多く提供されており、これにはデータ、html 形式のヘルプファイル、TEX 形式のファイル[*5]、pdf 形式のマニュアル、R の関数ファイル等が含まれています。パッケージのインストールについては Mac OS X については付録 C.1.2、Windows については付録 C.2.2 参照。

例題 1.5. R のパッケージ car のデータ Ericksen を考える。ここで、car は Companion to Applied Regression の頭文字です。 ■

入力 (出力結果は省略)

```
> library(car)      # ライブラリ car の組み込み。
> data(Ericksen)    # データ Ericksen の組み込む。
> help(Ericksen)    # データ Ericksen の説明表示。
> Ericksen          # データ Ericksen の内容表示。
```

help(Ericksen) から分かるように、66 行 9 列からなるデータで、各列の名前は names(Ericksen) により表示されます。

入力と出力結果)

```
> names(Ericksen)                    # データ Ericksen の列名表示。
[1] ''minority''     ''crime''        ''poverty''      ''language''
[5] ''highschool''   ''housing''      ''city''         ''conventional''
[9] ''undercount''
```

全米 50 州と 16 の大都市の 66 カ所に関する調査項目です。各項目が何を調査したものか、どの種類のデータかを確認しなさい。

例題 1.6. データ co2 を考える。これは、Hawaii, Mauna Loa の 1959 年から 1997 年までの毎月の二酸化炭素濃度の 468 個の観測値で、時間と共に観測されたデータです。時系列データと呼ばれます。■

[*5] TEX については http://www.biwako.shiga-u.ac.jp/sensei/kumazawa/tex.html 参照。R と同様に TEX もフリーウェアでさまざまな OS で利用でき、本稿も TEX によって作成されています。

入力

```
> data(co2)                # データ co2 の組み込み。
> ?co2                     # データ co2 の説明表示 (省略)。
> co2                      # データ co2 の内容表示 (省略)。
> plot(decompose(co2))     # データ co2 の時系列解析。図 1.3 参照。
```

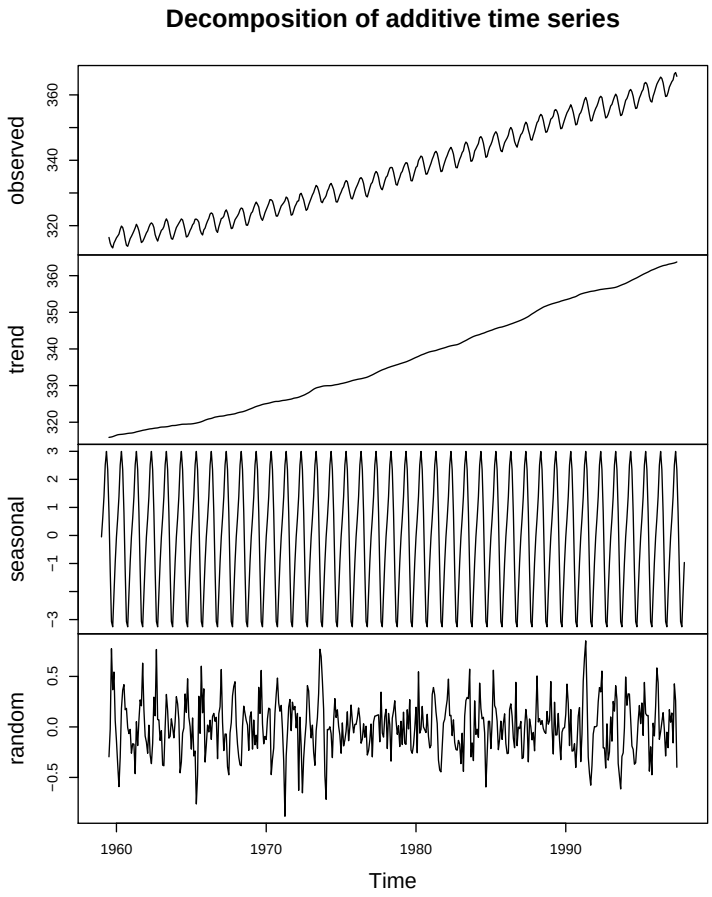


図 1.3　データ co2 の時系列解析

横軸は四つの図に共通な時間軸で、最上部の図 (observed) は観測された 468 個の二酸化炭素濃度のプロットで、その下の三つの図は時系列解析によって計算されたものである。第二の図 (trend) は時系列解析により計算された全体の傾向を表しており、図 (seasonal) は季節変動を示している。観測地が北半球にあるので、化石燃料の消費により二酸化炭素が冬に増加し、夏に減少することにより、季節変動はこの増減を示している。第四の図 (random) は時系列解析で導かれた傾向と季節変動以外の偶然変動を表している。

Tips library と data について

入力

```
> search()                  # ライブラリ一覧。
> library()                 # ライブラリ詳細。
> data()                    # データ一覧。
> help(package="car")
> data(package="car")
```

search() により、インストールされているパッケージの一覧が表示されます。library() により、パッケージの一覧詳細が表示されます。また、data() により、利用できるデータの一覧が表示されます。パッケージのインストールについては 91 ページ参照。help(package="car") とすることで、パッケージの内容と利用できるデータセットが表示されます。data(package="car") によりパッケージ car に含まれるデータの一覧が表示されます。

1.3　データの図式化

前節でソフトウェア R にはデータ解析のためのグラフィカルな機能とデータが備わっていることを見た。ここでは、他の記述統計関数と併せて説明を行う。

1.3.1　樹葉図

樹葉図(stem and leaf plot) とは、ヒストグラムでは階級に入っているデータについての情報が欠落するので、この点を補ったもので、データの概数を横軸にとり、これを樹と見なし、端数を葉と考え、描いたものである。図として描画されなく、コンソールへの出力となる。

右の例では標準正規乱数 rnorm を用いているので、実行する度に結果が異なってきます[*6]。

入力と出力結果

```
> help(stem)                    # 関数 stem の説明表示。
> x<- rnorm(100)                # 100 個の標準正規乱数。
> stem(x)                       # データ x の樹葉図の作成。
 The decimal point is at the |
 -1 | 999887
 -1 | 4433321100
 -0 | 88888887765555
 -0 | 44333333322222110
  0 | 000111111112333344
  0 | 5566666778899999
  1 | 000012222344
  1 | 6689
  2 | 4
  2 | 59
```

1.3.2　度数分布図

度数分布図(frequency histogram) とは、異なる属性ごとに観測されたデータを棒グラフの形で与えたものである。

入力と出力結果

```
> help(barplot)                 # 関数 barplot の説明表示。
> data(VADeaths)                # データ VADeaths の組み込み。
> VADeaths                      # データ VADeaths の内容表示。
> barplot(VADeaths, beside=TRUE)  # 度数分布図の作成。図 1.4(a) 参照。
```

[*6] 同一の乱数列を再度発生させたいときは RNGkind("Mersenne-Twister",normal.kind="Inversion");set.seed(1) を用いて乱数発生の初期値を固定させることが出来ます。

(a) 度数分布図　　(b) 積み上げ棒グラフ

図 1.4　データ VADeaths

1.3.3　積み上げ棒グラフ

異なる属性ごとに観測され、表の形で与えられたデータを積み重ね棒グラフの形で与えたものである。

入力と出力結果

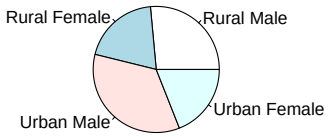
```
> barplot(VADeaths)    # 積み上げ棒グラフ。図 1.4(b) 参照。
```

1.3.4　円グラフ

円全体を 100% として，各属性の割合に従って順に区切り線を引き，塗り分けた図を**円グラフ**(pie chart) と呼ぶ。

入力と出力結果は右図

```
> t<- VADeaths[1,]    # データの第 1 行だけを用いる。
> pie(t)              # 円グラフの作成。
```



1.3.5　ヒストグラム

例題 3 で見たように、データ x に対して、hist(x) とするだけで**ヒストグラム**(histogram) が作成されます。ヒストグラムを描く際、階級数、階級幅等を決定する必要があるが、hist(x) とするだけで描画できる。詳細の設定に関しては?hist で現れる項目を設定することになる。ヒストグラムは 8 ページ挙げるパレート図と並び、品質管理の七つ道具の一つとなっている[*7]。

入力と出力結果

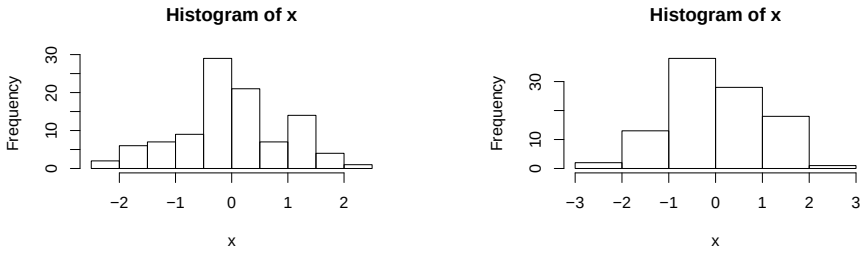
```
> help(hist)                    # 関数 hist の説明表示。
> x<- rnorm(100)                # 100 個の標準正規乱数を発生。
> par(mfrow=c(1,2))             # グラフを 1*2 個で表示する。
> hist(x)                       # Sturges の公式による x のヒストグラム。図 1.5(a) 参照。
> hist(x, breaks=seq(-3,3,1))   # 階級区間を-3 から 3 まで 1 刻みにしたヒストグラム。図 1.5(b) 参照。
> par(mfrow=c(1,1))             # グラフを 1*1 個に戻す。
```

[*7] 他の五つは、特性要因図、グラフ・管理図、散布図、チェックシート、層別で、第 8 章参照。

上のコマンド `hist(x, breaks=seq(-3,3,1))` において、`seq(-3,3,1)` は −3 から 1 刻みで 3 までの数列を表している。このコマンドにより、データセット `x` のヒストグラムが階級 $-3, -2, \cdots, 2, 3$ に基づき作成されることになる。また、ヒストグラムのヘルプファイル`?hist` には Sturges という名前があり、これはデータ数を n としたとき、階級数を

$$\text{階級数} = 1 + \log_2 n \fallingdotseq 1 + 3.321928 \times \log 10n$$

によって求める公式で、Excel の分析ツールにも組み込まれているスタージェス (Sturges) の公式です。



(a) デフォルトのヒストグラム　　(b) 階級区間を指定したもの

図 1.5　正規乱数のヒストグラム

1.3.6　パレート図

パレート (Pareto) の法則は、「重要なものは僅かしかない」ということを図式化したものです。**パレート図**(Pareto chart) は、分類項目の度数の高いものを高い順に並べ、縦軸の左軸に度数、右軸に累積百分率の目盛りを配したものである。パッケージ `qcc` を組み込むことでパレート図が作図されます[*8]。

入力と出力結果

```
> library(qcc)                                  # パッケージ qcc の組み込む。
> x<- c(80, 27, 66, 94, 33)                     # データ x の作成。
> names(x) <- c("A", "B", "C", "D", "E")        # x に名前を付与。
> x                                             # x の内容確認
 A  B  C  D  E
80 27 66 94 33
> pareto.chart(x, ylab = "Error frequency")     # y 軸に名前を付けてパレート図を作成。図 1.6 参照。
Pareto chart analysis for x
    Frequency Cum.Freq. Percentage Cum.Percent.
  D        94        94   31.33333     31.33333
  A        80       174   26.66667     58.00000
  C        66       240   22.00000     80.00000
  E        33       273   11.00000     91.00000
  B        27       300    9.00000    100.00000
```

[*8] パッケージについては 91 ページ参照。qcc は Quality Control Charts の頭文字。

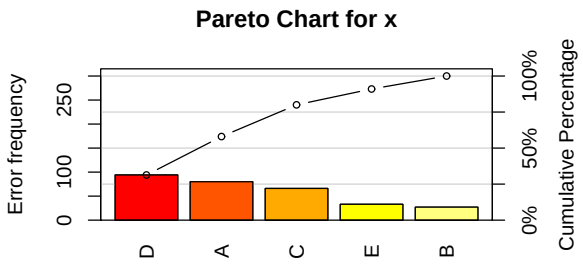


図 1.6　パレート図

1.3.7　正規確率プロット

正規確率プロット(normal quantile-quantile plot) とは、データが正規分布に従っているかを見るための図である。関数 `qqnorm` により、データの正規確率プロットが描かれ、関数 `qqline` により正規確率プロット上に第 1 四分位点と第 3 四分位点を結んだ直線が引かれる。

入力：出力結果は図 1.7

```
> help(qqnorm)          # 関数 qqnorm の説明表示。
> help(qqline)          # 関数 qqline の説明表示。
> x<- rnorm(100)        # 正規乱数を 100 個発生させ、x に代入。
> y<- runif(100)        # 一様乱数を 100 個発生させ、y に代入。
> par(mfrow=c(1,2))     # グラフを 1*2 個で表示する。
> qqnorm(x)             # x の正規確率プロット。
> qqline(x)             # x の qqline。
> qqnorm(y)             # y の正規確率プロット。
> qqline(y)             # y の qqline。
> par(mfrow=c(1,1))     # グラフを 1*1 個に戻す。
```

図 1.7 の左のものは 100 個の正規確率乱数に対するもので、右の図は 100 個の一様乱数に対して描かれた正規確率プロットと直線である。

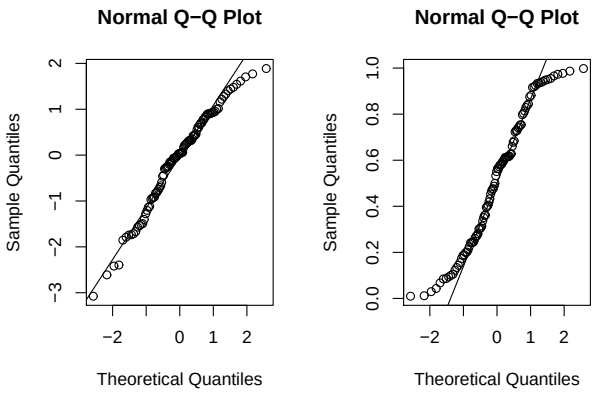


図 1.7　正規確率プロット

1.3.8　箱ひげ図

データを小さい方から順番に並べて、真中の値を**中央値**(median) といい、小さい方から 25% の位置にある値を**第 1 四分位点**(first quartile)、大きいほうから 25% の位置にある値を**第 3 四分位点**(third quartile) という[*9]。縦軸に変数値をとって、第 1 四分位点を下に、第 3 四分位点を上にした箱を書き、中央値の位置にも線を引いく。さらに第 1 四分位点と第 3 四分位点の差 (四分位範囲) を 1.5 倍した線分をヒゲとして第 1 四分位点の下と第 3 四分位点の上に伸ばし、ヒゲの先より外れた値を外れ値として○を描いた図を箱ひげ図 (boxplot) という。

入力：出力結果は図 1.8

```
> help(boxplot)          # 関数 boxplot の説明表示。
> par(mfrow=c(1,2))      # グラフを 1*2 個で表示する。
> boxplot(rnorm(100))    # 100 個の標準正規乱数の箱ひげ図。
> boxplot(rnorm(100))    # 別の 100 個の標準正規乱数の箱ひげ図。
> par(mfrow=c(1,1))      # グラフを 1*1 個に戻す。
```

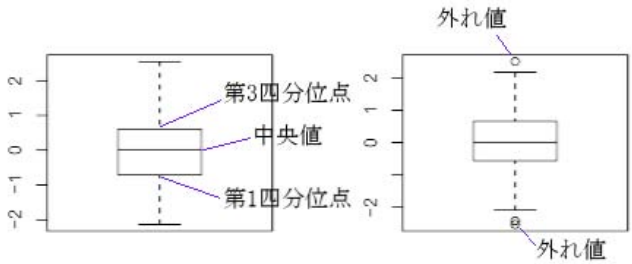


図 1.8　100 個の正規乱数の箱ひげ図

R の基本データとして組み込まれている、アヤメのデータ `iris` を考える。このデータは 3 種類のアヤメ (`setosa`、`versicolor`、`virginica`) について、ガク片の長さ・幅、花片の長さ・幅を測定した、各種について 50 個、計 150 個のデータである。データの列名は `Sepal.Length`、`Sepal.Width`、`Petal.Length`、`Petal.Width`、`Spieces` である。

入力：出力結果は図 1.9

```
> data(iris)                              # データ iris の組み込み。
> help(iris)                              # データ iris の説明表示。
> boxplot(Sepal.Length Species,data=iris)  # 種ごとの箱ひげ図。
```

`plot(Sepal.Length Species,data=iris)` としても同じ箱ひげ図が得られます。

[*9] 中央値、第 1 四分位点、第 3 四分位点については 24 ページと 25 ページ参照。

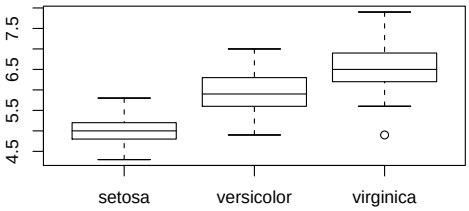


図 1.9　アヤメの種ごとの箱ひげ図

1.3.9　散布図

一つの個体に対して測定された 2 つの量的データの値を描いた図を**散布図**(scatter diagram) という。アヤメのデータ `iris` を考える。

入力：出力結果は図 1.10

```
> data(iris)    # データ iris の組み込み。
> help(iris)    # データ iris の説明表示。
> plot(iris)    # データ iris のすべての項目についての散布図。
```

`pairs(iris)` としても同じ散布図が得られます。

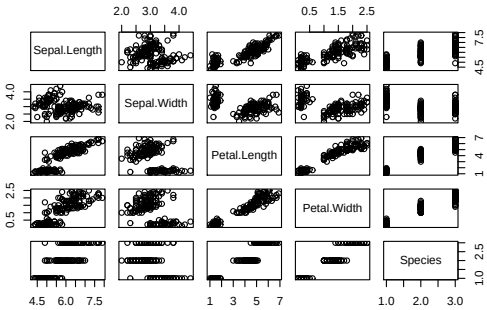


図 1.10　アヤメの散布図

関数 `plot` と `pairs` の Examples にはデータ記述に役立つものが多くのコマンドが多く記述されています。
次に、例題 1.2 で考えたデータ `cars` を考える。

入力：出力結果は図 1.11

```
> data(cars)                          # データ cars の組み込み。
> help(cars)                          # データ cars の説明表示。
> plot(cars)                          # 散布図の作成。
> abline(lm(dist   speed, data=cars))  # 回帰直線の作成。
```

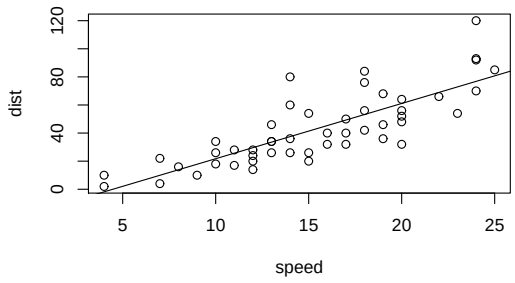


図 1.11　データ cars の散布図と回帰直線

関数 lm(dist　speed ,data=cars) は dist に対する speed の回帰直線の推定値を与え、abline はこの回帰直線式を描くというものである。コマンド lm(dist　speed ,data=cars) を入力することで、直線回帰の傾きと切片の推定値が得られる。また、summary(lm(dist　speed, data=cars)) と入力することで、直線回帰についての検定結果が得られます[*10]。

1.3.10　モザイクプロット

測定された分割表データを面積として表した図を**モザイクプロット**(mosaic plot) という。

入力：出力結果は図 1.12

```
> data(Titanic)          # データ Titanic の組み込み。
> help(Titanic)          # データ Titanic の説明表示。
> mosaicplot(Titanic)    # データ Titanic のモザイクプロット。
```

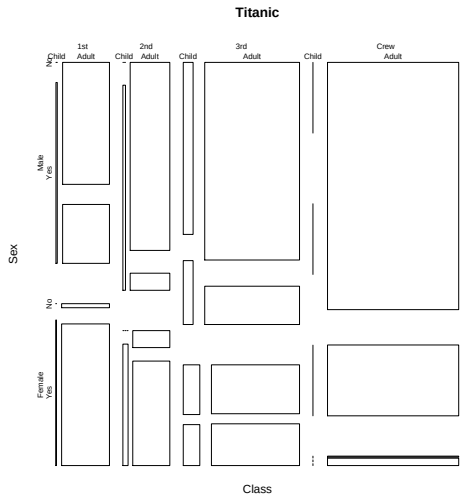


図 1.12　データ Titanic のモザイクプロット

[*10] 68 ページ参照。

Tips example について

help を見る必要はないが、関数の使用例を見たいときに example を用います。このとき、作図例が複数個あるとき、直ぐに次の作図例の出力に移ってしまうので par(ask=T) を入れておくことで、次の例に移ってもいいかの問い合わせがあり、例題が見やすいくなります。TRUE は T としてもよく、par(ask=F) を入力すると、次の描画についての問いがなくなります。F は False とできます。

入力

```
> par(ask=T)
> example(plot)    # plot の例を表示。
```

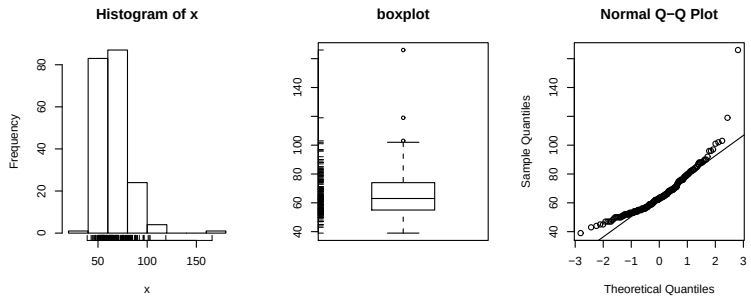
問題

問題 1.1.

関数 pairs の例題 example(pairs) を実行せよ。

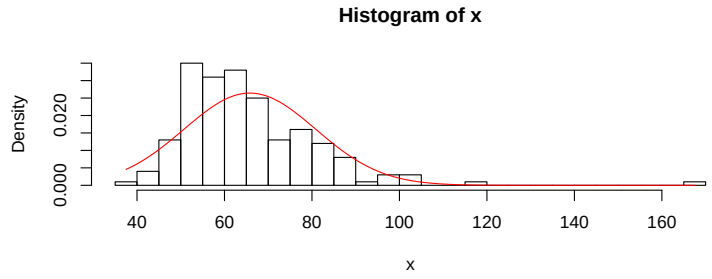
問題 1.2.

適当なデータに対して、ヒストグラム、箱ひげ図と正規確率プロットを描け。



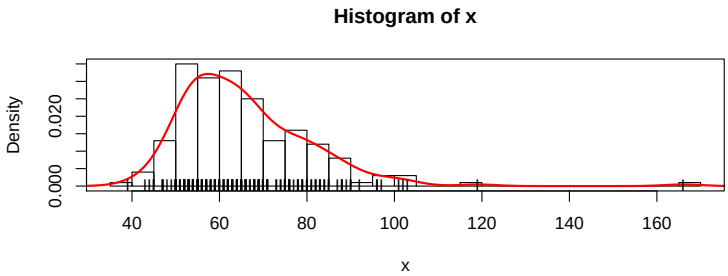
問題 1.3.

適当なデータに対して、ヒストグラムを描き、そのデータの平均と分散から決定される正規密度関数を描け。



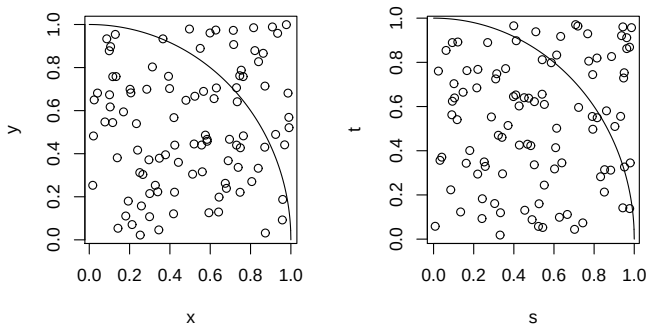
問題 1.4.

適当なデータに対して、ヒストグラムを描き、そのデータに対する密度関数を描け。



問題 1.5.

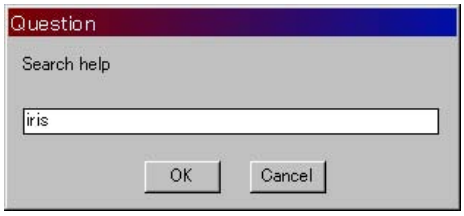
一様乱数を 2 組 100 個発生させ、二乗和が 1 より小さくなるものを数える。この値を 4 倍し、100 で割ったものは円周率 π の近似値と考えられる。この近似値を出力するプログラムを書け[*11]。



左図は 2 組 100 個一様乱数を発生させ、プロットしたもので、近似値 3.28 を得、右図は 2 組 10,000 個一様乱数を発生させ、プロットしたもので、近似値 3.04 を得た。発生させる乱数の個数を増加させることで円周率に近い値を得ることが出来る。

Tips `help` について — 2

R のメニューバーの help にある `Search help` を選ぶと、検索したい用語を入力するダイアログ



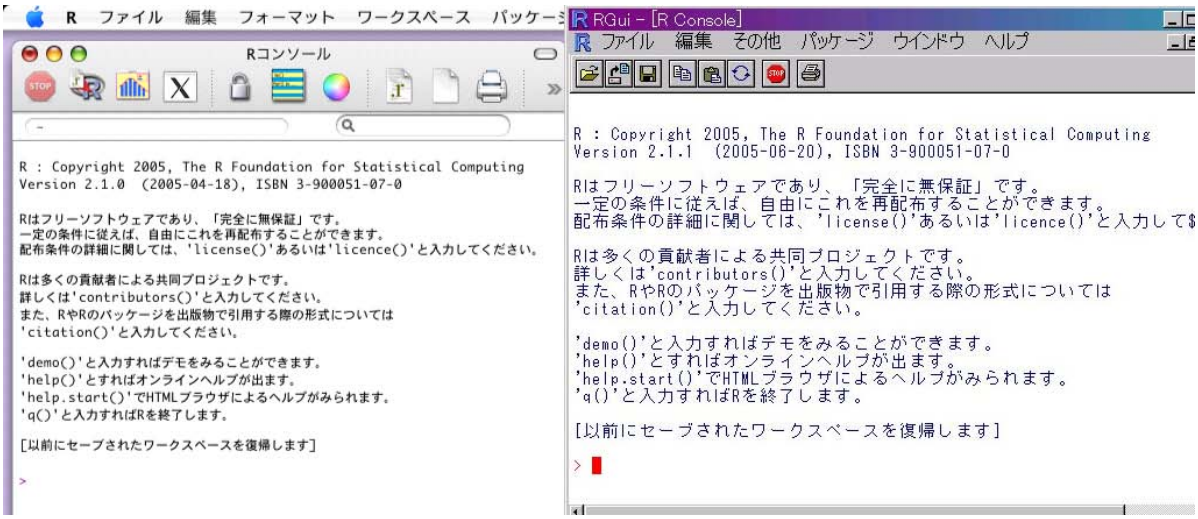
が現れます。日本語対応でないので、英語の用語を入れ、“OK” をクリックすると説明が現れます。これは、コンソールから `help.search("iris")` と入力しても同じ説明が得られます。

[*11] 乱数の初期設定は `set.seed` を用います。

第 2 章　データの入力と計算

2.1　ソフトウェア R の起動と終了

ソフトウェア R のアイコンをクリックすることで、R が起動します。左が Mac OS X 版で、右が Windows 版の起動時の画面です。



R を終了するには下記のいずれかを行います。

- コンソールに q() または、quit() を入力し、Enter キー↵を押す。
- ウィンドウの終了ボタンをクリックする。
- Mac のときはメニューの R から Quit を選ぶ。Windows のときはメニューの File から Exit を選ぶ。
- Mac のときはコマンドキーを押しながら Q を押す。Windows のときは Alt キーを押しながら F4 キーを押す。どちらの操作も他のソフト終了時に利用できます。

終了を選択した後、作業内容を保存するかのダイアログが出るので、履歴を残したいときは、Mac OS X 版では “save”、Windows 版では “はい (Y)” を選択します。

2.2　データの入力

コンソールから数値を含む式を入力し、対話形式で結果を得ることができます。

データは、論理数、実数、複素数、文字列のいずれかで、これに基づき、ベクトル、表形式、時系列等のオブジェクトが構成される。論理数は正しい (真) か誤りである (偽) を示す TRUE(または T) または FALSE(または F) という値である。実数は日頃用いているのと同じ表記法に従い、複素数は 4+6i のように表す。

入力と出力結果

```
> 1+2          # 1+2 とタイプし、Enter キー↵を押す。
[1] 3          # 結果の 3 が出力される。
```

文字列は文字を一対の二重引用符 (") で囲んだもので、アンケートの回答などに返答がない場合には欠測値をしめす NA (Not Available) を用いることになります。

入力と出力結果

```
> month.name                                          # R に組み込まれている月の名前。
 [1] "January"   "February"  "March"     "April"
 [5] "May"       "June"      "July"      "August"
 [9] "September" "October"   "November"  "December"
> x<- NA                                              # x に NA を代入。
> x                                                   # x の内容表示。
[1] NA
```

2.2.1　代入

すでに<-を用いて、<-の左にある変数に<-の右の値が代入(付値) されることをみた。R では<-と同等なものとして、

-> と =

が用意されています。

入力と出力結果

```
> x<- pi   # 円周率を x に付値。
> pi -> x
> x = pi
```

円周率 pi は R に組み込まれた予約語で

break　else　for　function　if　in　next　repeat　return　while

と同様に決められた役割を持ちます。

また、右図のようなに 2 つのオブジェクトに同じ値を同時に付値することが出来ます。

入力と出力結果

```
> a<- b<- 0   # a、b に 0 を代入。
> a
[1] 0
> b
[1] 0
```

入力と出力結果

```
> a<- 0      # a にゼロを代入。
> a< 0       # a が 0 より小さいか判定。
[1] FALSE    # a の値は 0 でるので FALSE(偽)。
> b==0       # b の値は 0 か判定。
[1] TRUE     # b の値は 0 であるから TURE(真)。
```

注意が必要なのは<と>が比較のためにも利用されているということです。付値<-を用いるときは、<-␣とタイプすることで、

x<-␣2　　x に 2 を付値。
x<␣-2　　x と-2 を比較。

の区別をはっきりさせることが出来ます。ここで、␣は半角の空白を表しています。

2.2.2　ベクトル

関数 c を用いて数値、文字列のベクトルを定義します。

入力と出力結果

```
> x<- c(1,2,3,4)                    # x に (1,2,3,4) を代入。
> x                                 # x の表示。
[1] 1 2 3 4
> y<- c("day", "evening","night")    # y に"day", "evening","night"を代入。
> y
[1] "day"     "evening" "night"
```

2.2.3　行列と表

縦方向と横方向に数値を並んだものを作ることを考える。すでに 3 ページの cars や Titanic でこのような形のデータ形式を見ている。

例題 2.1. 自動車免許の所有を調査したところ下記のデータを得た。license という変数に値を代入することを考える。■

	なし	原付	自動車
男	18	3	20
女	20	10	15

関数 data.frame を用いて、右のように定義することが出来る。オブジェクト non、bike、car に列の数値を代入し、data.frame を用いて、表の形にしています。

入力と出力結果

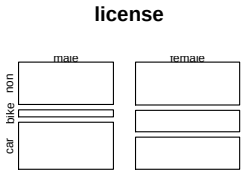
```
> non<- c(18,20)                                  # non の付値。
> bike<- c(3,10)                                  # bike の付値。
> car<- c(20,15)                                  # car の付値。
> license<- data.frame(non, bike, car)
> license
  non bike car
1  18    3  20
2  20   10  15
> rownames(license)<- c("male","female")   # 行名を付値。
> license
       non bike car
male    18    3  20
female  20   10  15
```

入力と出力結果

```
> matrix(c(18,20,3,10,20,15),nrow=2)-> license
> rownames(license)<- c("male","female")        # 行名。
> colnames(license)<- c("non","bike","car")     # 列名。
```

別の方法として、関数 matrix を用いた左のものがある。

12 ページの関数 mosaicplot を使うと右図を得る。例題 6.4(57 ページ) において適合度検定が適用され、男女間に差がないという仮説の下でこのようなデータが起こる確率が 0.11 であることが示される。



このデータ license に対して、横合計と縦合計を追加した集計表を作成する。横合計 (行和) と縦合計 (列和) を計算する関数として関数 margin.table がある。関数 margin.table(,1) により行和 (横合計) が求まり、margin.table(,2) により列和 (縦合計) が求まります。

入力と出力結果

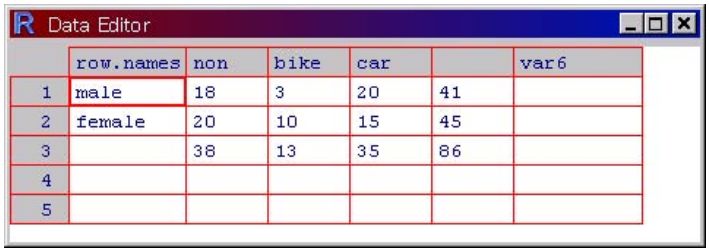
```
> margin.table(license,1)
  male female
    41     45
> margin.table(license,2)
 non bike  car
  38   13   35
```

入力と出力結果

```
> cbind(license, margin.table(license,1)) ->license
> rbind(license, margin.table(license,2)) ->license
> license
       non bike car
male    18    3  20 41
female  20   10  15 45
         38   13  35 86
```

さらに、関数 cbind と rbind を組み合わせて集計表が得られる。関数 cbind(x,y) は列 (column) に関する操作で同じ行数のものを結合するもので、rbind(x,y) は行 (row) に関する操作で同じ列数のものを結合するものある。

データや表形式のデータを修正する際便利な関数として、fix がある。上のデータ license を修正したいとき、fix(license) とすると、Data Editor



が起動し、表計算ソフトのセルの修正のように操作を行うことが出来、修正した値は license に保存されます。これはメニューの edit から Data Editor を選択し、オブジェクト名を入力しても同様な修正を行うことが出来ます。

数列を作る際、便利なものとして関数:があり、1 刻みで数列を作成します。1:10 により、1 から 10 までの数列が作られ、-1:5 により、-1 から 5 までの数列が作られます。

入力と出力結果

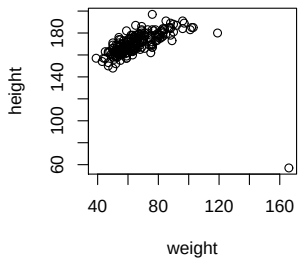
```
> 1:10
 [1]  1  2  3  4  5  6  7  8  9 10
> -1:5
[1] -1  0  1  2  3  4  5
```

入力と出力結果

```
> seq(1,10, by=3)          # 1 から 10 まで 3 刻みの数列。
 [1]  1  4  7 10
> seq(10,-3 ,by=-2)        # 10 から-3 まで-2 刻みの数列。
[1] 10  8  6  4  2  0 -2
```

関数 seq の by によって数列の増分を指定することが出来ます。

ライブラリ car のデータ Davis を考える[*1]。このデータは 200 人の実際の身長・体重と申告された身長・体重に関するものである。実際の身長と体重の散布図を描く[*2]と右図のようになり、身長の単位は cm、体重の単位は kg なので、図の右隅の点を表すデータは間違って入力されたものと考えられる。このデータ番号はコマンド identify(weight, height) を用いることで散布図内に表示することが出来ます。コマンド入力後、散布図の点を左クリックすることでデータ番号が散布図に書き加えられ、右クリックをすることでデータ番号の表示を終了することが出来ます。このコマンドの終了時にはコンソールにもデータ番号が出力されます。



[*1] library(car); Data(Davis)
[*2] plot(weight, height, data=Davis)

実際、データ番号 12 のデータは

```
> Davis[12,]
   sex weight height repwt repht
12   F    166     57    56   163
```

となっています。この 12 番目のデータは数値自体も正しいものが入力されていると考えずに取り除くには、– 記号を用いて

```
> Davis1<- Davis[-12,]
```

とします。新たに出来たオブジェクト Davis1 の内容表示をすると 12 番目のデータが削除されていることが分かります。データ番号 12 の実際の身長・体重のデータを入れ替えるには、関数 fix を用い、

```
> Davis2<- Davis
> Davis2<- fix(Davis2)
```

により、修正したオブジェクト Davis2 が得られます。ここで、上の 2 行のコマンドを Davis2<- fix(Davis) とすると、Davis のデータも修正されるので注意が必要です。

実際の身長は 2 列目、体重は 3 列目なので、下記のように修正することも出来ます。

```
> Davis2<- Davis
> Davis2[12,2]<- 57
> Davis2[12,3]<- 166
> Davis2[12,]
   sex weight height repwt repht
12   F     57    166    56   163
```

2.2.4　時系列データ

12 ページにおいてデータ co2 の時系列解析を図示した。ここではその時系列データの作成を考える。

例題 2.2. 三週間の受信メール数を集計したら、以下のようになった。時系列データとしてデータを構成せよ。■

	日	月	火	水	木	金	土
第 1 週	37	62	89	64	68	32	27
第 2 週	34	63	83	63	65	34	23
第 3 週	41	67	82	61	65	36	24

入力と出力結果は図 2.1

```
> x<- c(37,62,89,64,68,32,27,        # 第一週目の受信メール数。
      34,63,83,63,65,34,23,          # 第二週目の受信メール数。
      41,67,82,61,65,36,24)          # 第三週目の受信メール数。
> ts(x, frequency=7)-> y             # x を頻度 7 の時系列データに変換。
> plot(decompose(y))                 # データ y の時系列解析の図式化。
```

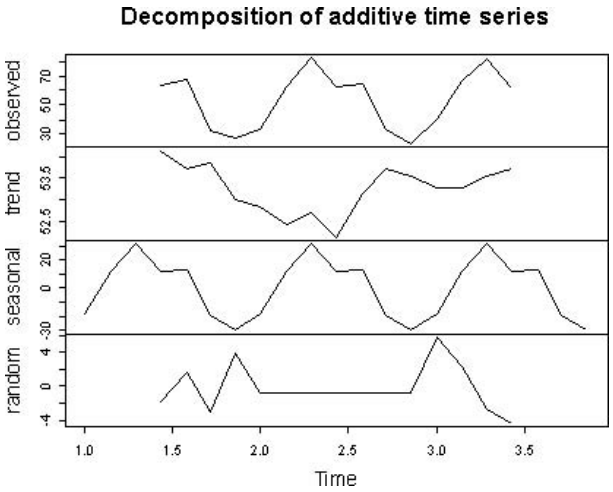


図 2.1　時系列解析

2.2.5　csv ファイルの読み込み

csv ファイル (comma separated value file) は、数値データ、文字データがカンマを用いて区切られているファイルのことで、Excel がインストールされているパソコンでは Excel のファイルとして扱われるが、テキストファイルなので、エディタ[*3]を用いて編集することが出来ます。

`setwd`(**s**et **w**orking **d**irectory) によってデータが保存されている directory「`F:\data_A`」を指定し、`read.csv` によって csv ファイルを読み込みます。SPSS や SAS などの他の統計ソフトのデータを直接読み込むには `foregin` というパッケージがあります。

入力と出力結果

```
> setwd("F:\data_A")
> read.csv("csv01.csv")
   X.1 x A X18.33
1   -1 x B  10.21
2   -1 x C  -2.00
3   -1 x D -14.02
4   -1 y A -21.37
```

上の手法を用いると多量のデータを含むファイルを読み込むのに時間がかかることがあります。データのファイルを R で扱うことの出来る形にし、拡張子を `r` にして保存し、このファイルを R のコンソールにドラッグ&ドロップすると読み込む時間を短縮することができます。この方法はデータのみならず実行形式のファイルにも適応しています。

2.2.6　ファイルへの書き出し

通常の操作において出力先はコンソール、グラフィックまたはヘルプ画面である。多量の出力結果があるとき、エディタを用いてコンソールの内容をコピー・貼り付けするのは面倒である。テキスト形式のファイル出力として関数 `sink` がある。

入力

```
> setwd("F:\data_A")    # ファイルを書き出すフォルダーに移動。
> sink("test.txt")      # ファイル"test.txt"に関数 sink() があるまで書き出す。
> a                     # a の内容はコンソールに表示されない。
> sink()                # 書き出しの終了。
```

[*3] 105 ページ参照。

2.3　データの計算

オブジェクト間の計算としては四則演算に加え、下記のものがある。

記号	+	−	/	*	∧	%%	%/%
意味	加算	減算	乗算	除算	べき乗	剰余	整数商

入力と出力結果

```
> 1*2+3/4
[1] 2.75
> 3%%5        # 余り。
[1] 3
> 3%/%5       # 商の整数部。
[1] 0
```

入力と出力結果

```
> 6%%5            # 余り。
[1] 1
> 6%/%5           # 商の整数部。
[1] 1
> 2^52            # べき乗。
[1] 4.5036e+15
```

2.3.1　初等関数

通常用いる代表的な関数と同様な関数が定義されています。これ以外にも初等関数が用意されています。

記号	sin	cos	tan	sinh	cosh	tanh
意味	sin	cos	tan	sinh	cosh	tanh
記号	asin	acos	atan	$\sinh^{-1}$	$\cosh^{-1}$	$\tanh^{-1}$
意味	$\mathbf{Sin}^{-1}$	$\mathbf{Cos}^{-1}$	$\mathbf{Tan}^{-1}$	$\sinh^{-1}$	$\cosh^{-1}$	$\tanh^{-1}$
記号	log	log10	log2	exp	sqrt	
意味	log	$\log_{10}$	$\log_2$	exp	$\sqrt{\ }$	

入力と出力結果

```
> sin(0);sin(pi/2);sin(pi)
[1] 0
[1] 1
[1] 1.224606e-16
> cos(0);cos(pi/2);cos(pi)
[1] 1
[1] 6.123032e-17
[1] -1
> x<- pi/3
> sin(x)^2+cos(x)^2
[1] 1
```

関数 sin、cos、tan は弧度法でなくラジアンで測ったものを用いています。つまり、0 度を 0 とし、180 度を π として測定したものを用いています。

0°	···	30°	···	60°	···	180°	···	360°
0	···	$\frac{\pi}{6}$	···	$\frac{\pi}{3}$	···	π	···	2π

セミコロン (;) によって複数行のコマンドを一行にまとめることが出来ます。

2.4　オブジェクトの一覧と削除

データ解析を行った結果、定義したオブジェクト (変数) の一覧が必要になることがある。このとき、一覧表示には ls() を用います。何もオブジェクトが保存されていないときは、character(0) が出力されます。

データセットやライブラリの例題を実行したとき、上手く動かないことがあります。これは例題内で関数 attach が用いられ、例題内のオブジェクト名といままで用いてきたオブジェクト名が重複したことによるものです。このときは、関数 rm により重複している変数を削除します。また、すべてのオブジェクトを削除したいときは rm(list=ls()) を入力します。

入力と出力結果

```
> ls()                    # オブジェクトの一覧。
[1] "sample"
> example(qcc)            # ライブラリ qcc の例題。
qcc> data(pistonrings)
qcc> attach(pistonrings)
qcc> diameter <- qcc.groups(diameter, sample)
Error in qcc.groups(diameter, sample) : data and sample must be vectors of equal length
```

オブジェクト sample がすでにあるとき、example(qcc) を実行し、pistonrings の sample と重複しているためにエラーが起きている。定義した sample を他のオブジェクトに移動し、再度、example(qcc) を実行することで qcc の例題を表示することが出来ます。

入力

```
> sample1<- sample   # sample を sample1 に付値。
> example(qcc)       # ライブラリ qcc の例題。
```

Tips　options について

入力と出力結果

```
> pi                    # π の表示。
[1] 3.141593
> options(digits=5)     # 表示桁数を 5 桁にする。
> pi
[1] 3.1416
```

通常の表示桁数は 7 であるが、digits を用いて変更できます。最大表示桁数は 22 桁です。

第 3 章　データの数値化

データを要約し数値化を考えるとき、代表的なものとして、位置の尺度、散らばりの尺度、データ間の関連、データの特性をみる量がある。

- 数値化
  - 位置の尺度
    - 算術平均　幾何平均　調和平均　重み付け平均
    - 中央値　最頻値　最大値　最小値
    - 下側ヒンジ　上側ヒンジ　第 1 四分位点　第 3 四分位点
  - 散らばりの尺度
    - 分散　標準偏差　平均偏差　範囲
    - ヒンジ散布度　四分位点範囲
  - データ間の関連
    - 相関係数　共分散
  - データの特性をみる量
    - 歪度　尖度　変動係数　Gini の係数

3.1　位置の尺度

3.1.1　算術平均

算術平均(arithmetic mean) は、データ $X_1, X_2, \cdots, X_n$ に対して

$$\overline{X} = \frac{1}{n}\{X_1 + X_2 + \cdots + X_n\} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

により定義される。特に、標本から計算されたことを示すのに**標本平均**(sample mean) と呼ぶことがあり、R では関数 `mean` で定義されています。

データ `cars` の `speed` の平均を計算すると

```
> data(cars); x<- cars$speed
> mean(x)
[1]}\ 15.4
```

となります。この節を通じて、`x` はこの値を用います。

3.1.2　幾何平均

幾何平均(geometric mean) は、正の値のデータ $X_1, X_2, \cdots, X_n$ に対して

$$\overline{X}_G = \sqrt[n]{X_1 X_2 \cdot X_n} = \sqrt[n]{\prod_{i=1}^{n} X_i}$$

によって定義される。平均物価上昇率、平均成長率の計算に用いられる。R には定義されていないのでコマンド `prod(x)^(1/length(x))` を用いることになります。

```
> prod(x)^(1/length(x))  # x の幾何平均
[1] 14.32501
```

3.1.3　調和平均

調和平均(harmonic mean) は、データ $X_1, X_2, \cdots, X_n$ に対して

$$\overline{X}_H = \frac{1}{\frac{1}{n}\left(\frac{1}{X_1} + \frac{1}{X_2} + \cdots + \frac{1}{X_n}\right)} = \frac{1}{\frac{1}{n}\sum_{i=1}^{n}\frac{1}{X_i}}$$

によって定義される。R には定義されていないのでコマンド `1/sum(1/x)*length(x)` を用いることになります。

```
> 1/sum(1/x)*length(x)   # x の調和平均
[1] 12.96153
```

3.1.4　重み付け平均

データ $X_1, X_2, \cdots, X_n$ を小さい順に並べたものを

$$X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$$

によって表し、これを順序統計量 (order statistics) とよぶ。このとき、**重み付け平均**(weighted mean) は、データ $X_1, X_2, \cdots, X_n$ と重み $c_1, c_2, \cdots, c_n$ に対して

$$\overline{X}_W = \frac{1}{\sum_{i=1}^{n} c_i}\{c_1X_{(1)} + c_2X_{(2)} + \cdots + c_nX_{(n)}\} = \sum_{i=1}^{n} c_iX_{(i)}$$

によって定義される。極値 (最小値や最大値) に重みを小さくして極値からの影響を少なくして計算するために用いられることがある。

```
> n<- length(x);wgt<- rep(1,n);wgt[1]<- wgt[n]<- 0
> sum(wgt*sort(x))/n
[1] 14.82
```

重み `wgt` の第一成分と第 `n` 成分をゼロとし、その他の重みを 1 として、重み付け平均を計算しています。

3.1.5　中央値

このとき、**中央値**(median) は、順序統計量 $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$ に対して、

$$\tilde{X} = \begin{cases} X_{(\frac{n+1}{2})} & (n \text{ が奇数のとき}) \\ \frac{1}{2}(X_{(\frac{n}{2})} + X_{(\frac{n}{2}+1)}) & (n \text{ が偶数のとき}) \end{cases}$$

によって定義され、R では `median` により計算されます。

```
> median(x)
[1] 15
```

3.1.6　最頻値

最頻値(mode) は、データの中で最も多くの頻度で発生した値であり、ヒストグラムを作成したときの山の一番高い階級の中央値として定義される。同じ高さの山があるときは山に対応したすべての値が最頻値となります。

3.1.7　最小値

最小値(minimum) は、データ $X_1, X_2, \cdots, X_n$ の最小値 (順序統計量 $X_{(1)}$) で、R では `min` により計算されます。

```
> min(x)
[1] 4
```

3.1.8　最大値

最大値(maximum) は、データ $X_1, X_2, \cdots, X_n$ の最大値 (順序統計量 $X_{(n)}$) で、関数 `max` によって計算されます。

```
> max(x)
[1] 25
```

3.1.9　下側ヒンジ

下側ヒンジ(lower hinge) は、データ $X_1, X_2, \cdots, X_n$ の中央値より小さな値の集まりの中央値として定義され、R では関数 `fivenum` の 2 番目の要素として与えられます。

```
> fivenum(x)
[1] 4 12 15 19 25
> fivenum(x)[2]
[1] 12
```

関数 `fivenum` はデータセットの最小値、下側ヒンジ、中央値、上側ヒンジ、最大値の五つの値を同時に出力します。

3.1.10　上側ヒンジ

上側ヒンジ(upper hinge) は、データ $X_1, X_2, \cdots, X_n$ の中央値より大きな値の集まりの中央値として定義され、R では関数 `fivenum` の 4 番目の要素として与えられます。

```
> fivenum(x)[4]
[1] 19
```

3.1.11　第 1 四分位点

第 1 四分位点(first quantile) は、データ $X_1, X_2, \cdots, X_n$ の順序統計量 $X_{(1)} \leq X_{(2)} \cdots \leq X_{(n)}$ の小さな方から 25% にある値として定義され、R では関数 `summary` の 2 番目の要素として与えられます。節 1.3.8 の `boxplot` において、次の第 3 四分位点と併せて用いられていました。

```
> summary(x)                                  # x の要約値。
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    4.0    12.0    15.0    15.4    19.0    25.0
> summary(x)[2]                               # x の第 1 四分位点。
1st Qu.
     12
```

関数 `summary` は最小値、第 1 四分位点、中央値、平均、第 3 四分位点、最大値を六つの値を同時に計算します。

3.1.12　第 3 四分位点

第 3 四分位点(third quantile) は、データ $X_1, X_2, \cdots, X_n$ の順序統計量 $X_{(1)} \leq X_{(2)} \cdots X_{(n)}$ の小さな方から 75% にある値として定義される。R では関数 `summary` の 5 番目の要素として与えられます。

```
> summary(x)[5]   # x の第 3 四分位点。
3rd Qu.
   19.0
```

3.2　散らばりの尺度

3.2.1　分散

分散(variance) は、データ $X_1, X_2, \cdots, X_n$ に対して、平均 $\overline{X}$ からの X_i の**偏差**(deviation)$X_i - \overline{X}$ の二乗和を $n-1$ で割ったもの

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

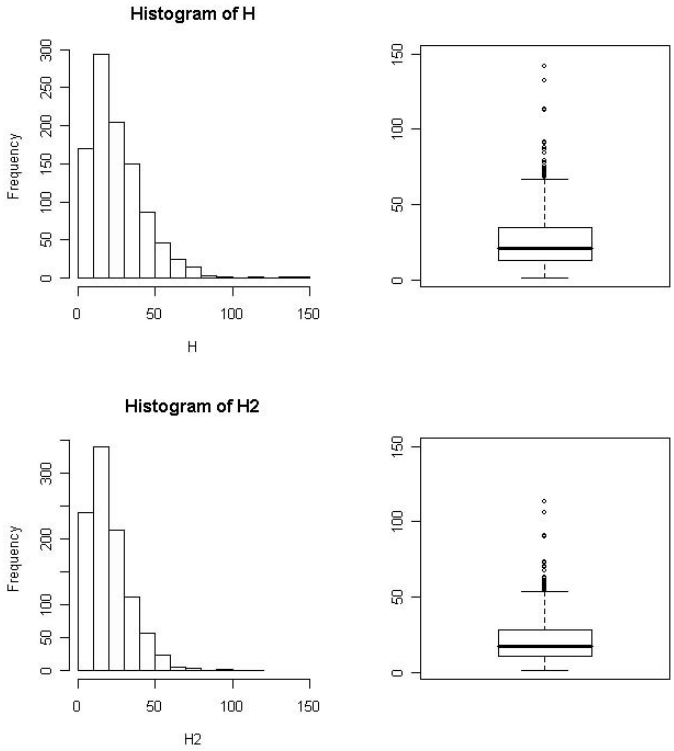
として定義され、R の関数 var はこの式を計算しています。この分散を不偏分散とも呼ぶ。特に、抽出された標本から計算された分散を**標本分散**(sample variance) と呼ぶ。標本数 n でなく、$n-1$ で割っていることに注意。

```
> var(x)
[1] 27.95918
```

平均=157、分散=25 の正規乱数を 5 個発生させ、この 5 つの標本からの分散と不偏分散を計算することを 1000 回繰り返すことを考える[*1]。

```
> H<-  as.vector(0)                    # 偏差平方和を 4 で割ったものを入れる。
> H2<-  as.vector(0)                   # 偏差平方和を 5 で割ったものを入れる。
> for(i  in  1:1000){                  # i の値を 1 から順に 1000 まで替えながら繰り返す。
> h<-  rnorm(5,mean=157,sd=5)          # 平均=157、分散=25 の正規乱数 5 個発生。
> H[i]<-  var(h)                       # 4 で割る。
> H2[i]<-  var(h)*4/5                  # 5 で割る。
> }                                    # 繰り返しの最後。
> par(mfrow=c(2,2))                    # 画面を 2*2 に分ける。
> hist(H)                              # H のヒストグラム。
> boxplot(H)                           # H の箱ひげ図。
> hist(H2)                             # H2 のヒストグラム。
> boxplot(H2)                          # H2 の箱ひげ図。
> par(mfrow=c(1,1))                    # 画面を元の 1*1 に戻す。
> summary(H)                           # H の要約値。
 Min. 1st Qu.  Median  Mean 3rd Qu.   Max.
 0.9581 12.7900 21.1000 25.6600 34.8500 141.7000
> summary(H2)                          # H2 の要約値。
 Min. 1st Qu.  Median  Mean 3rd Qu.   Max.
 0.7665 10.2400 16.8800 20.5300 27.8800  113.4000
```

上の要約値から H の平均は 25 に近く、H2 の平均は 20 に近いことが分かる。よって、データ数 −1 で割った方が真の分散に近いもの得られる。これは数学的に証明できる事柄である。



[*1] このシミュレーションは杉山 高一『統計学入門』(絢文社) 37 ページと同じである。

3.2.2　標準偏差

標準偏差(standard variance) は、分散の平方根

$$
\begin{aligned}
s &= \sqrt{s^2} \\
&= \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2}
\end{aligned}
$$

として定義され、R では sd によって計算されます。分散の測定単位が元の観測値の単位の二乗になるものに対して、標準偏差は観測値と同じ単位になるので、信頼区間、検定統計量、偏差値などの計算に利用される。

```
> sd(x)
[1] 5.287644
> sqrt(var(x))   # 分散の平方根
[1] 5.287644
```

3.2.3　平均偏差

平均偏差(mean deviation) は、平均 $\overline{X}$ から X_i の偏差 $X_i - \overline{X}$ の絶対値のを n で割ったもの

$$
MD = \frac{1}{n}\sum_{i=1}^{n}|X_i - \overline{X}|
$$

として定義される。R では `sum(abs(x-mean(x)))/length(x)` として計算されます。

```
> sum(abs(x-mean(x)))/length(x)
[1] 4.336
```

3.2.4　範囲

範囲(range) は、最大値 $X_{(n)}$ から引く最小値 $X_{(1)}$ を引いたもの

$$
R = X_{(n)} - X_{(1)}
$$

として定義され、データセット X に対して R の関数 `range(X)` は最小値と最大値の組を与えるので、`diff(range(X))` とすることで範囲が計算されます。

```
> range(x)
[1] 4 25
> diff(range(x))   # range(x) の差。
[1] 21
> max(x)-min(x)    # 最大値引く最小値。
[1] 21
```

3.2.5　ヒンジ散布度

ヒンジ散布度(hinge spread) は、上側ヒンジから下側ヒンジを引いたものとして定義される[*2]。

```
> f<- fivenum(x)
> f[4]-f[2]
[1] 7
```

3.2.6　四分位範囲

四分位範囲(interquantile range) は、第 3 四分位点から第 1 四分位点を引いたものとして定義され、R の関数 IQR によって与えられます。

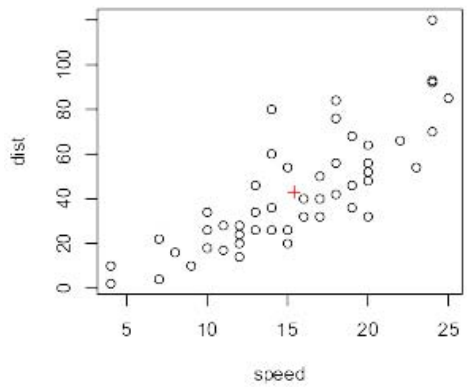
[*2] 25 ページ参照

```
> IQR(x)
[1] 7
```

コマンド `summary(x)[5]-summary(x)[2]` によっても四分位範囲が求まります。

3.3　データ間の関連

対になっているデータ (X_i, Y_i), $(i = 1, 2, \cdots, n)$ を考える。これは第 i 番目の個体に対して、体重と身長を測定したときのように個体 i について一組の観測値である。



3.3.1　共分散

データ (X_i, Y_i) の**共分散**(covariance) は

$$s_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})$$

で定義され、R では `cov` によって計算されます。ここで

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \qquad \overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

である。

```
> data(cars)                    # データ cars の組み込み。
> cov(carsspeed, carsdist)
[1] 109.9469
```

3.3.2　相関係数

データ (X_i, Y_i) の**相関係数**(correlation coefficient) は、

$$\rho = \frac{s_{XY}}{s_X s_Y}$$

により定義され、R では `cor` によって計算されます。ここで、s_{XY} はデータ (X_i, Y_i) の共分散で

$$s_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2, \quad s_Y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2$$

である。

```
> cor(carsspeed, carsdist)
[1] 0.8068949
```

3.4　データの特性をみる量

データ $X_1, X_2, \cdots, X_n$ に対して、原点のまわりの k **次モーメント**(moment) は

$$a_k = \sum_{i=1}^{n} X_i^k$$

として定義され、X_i の平均 $\overline{X}$ のまわりの k 次モーメントは

$$m_k = \sum_{i=1}^{n}(X_i - \overline{X})^k$$

として定義される。ここで、k は自然数である。モーメントの代わりに積率と呼ぶこともある。

3.4.1　歪度

歪度(わいど)(skew) とは、データ $X_1, X_2, \cdots, X_n$ の形成する分布の非対称性を測るもので、

$$S_K = \sqrt{n}\frac{\sum_{i=1}^{n}(X_i - \overline{X})^3}{\left\{\sum_{j=1}^{n}(X_j - \overline{X})^2\right\}^{3/2}}$$

によって定義され、分子は平均のまわりの 3 次モーメントであるから平均に対する非対称性を表し、0 に近いとき分布は対称であると考えたりします。歪度の大まかな性質は図 3.1 参照。

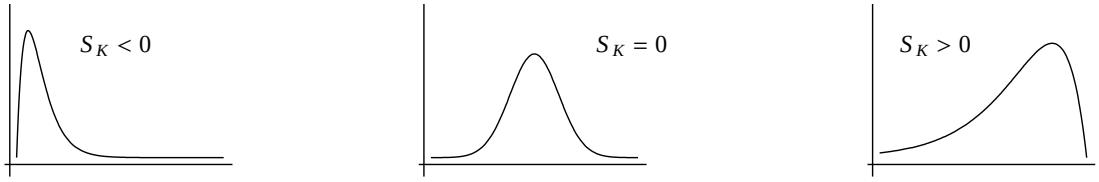


図 3.1　歪度

Excel の統計関数 SKEW は

$$\frac{n}{(n-1)(n-2)}\sum_{i=1}^{n}\left(\frac{X_i - \overline{X}}{s}\right)^3$$

で定義されています。ここで s は標準偏差で、データ数 n が 2 以下の時エラーを表示します。

R では歪度を求める関数が定義されていないので、自分で定義することになります。`function` により、新しい関数を定義します。`function` 内で新たに用いる変数は作業変数となり、関数の外の値とは区別されます。library“`e1071`"を組み込むと、歪度の計算に関して `skewness` という関数を利用することが出来ます[*3]。

関数 `skew` の定義

```
> skew<- function(x){                    # 関数 skew の定義。
> m3<- sum((x-mean(x))^3)/length(x)
> s3<- sqrt(var(x))^3
> m3/s3
> }
> skew(x)                                # skew の値の出力。
```

3.4.2　尖度

尖度(せんど)(kurtosis) とは、データ $X_1, X_2, \cdots, X_n$ の形成する分布の尖り度を測るもので

$$\kappa = \sqrt{n}\frac{\sum_{i=1}^{n}(X_i - \overline{X})^4}{\left\{\sum_{j=1}^{n}(X_j - \overline{X})^2\right\}^{2}} - 3$$

によって定義される。第一項の分子は平均のまわりの 4 次モーメントで、正規分布を用いると第一項の値が 3 であるので、3 と比較されています。尖度の性質については、図 3.2 参照。

[*3] ライブラリについては付録 C のライブラリのインストール参照。

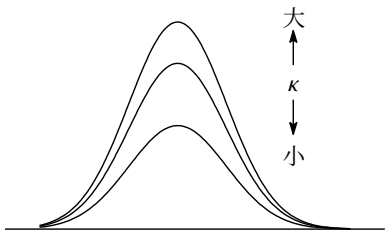


図 3.2　尖度

Excel の統計関数 KURT は

$$\frac{n(n+1)}{(n-1)(n-2)(n-3)}\sum_{i=1}^{n}\left(\frac{X_i-\overline{X}}{s}\right)^4-\frac{3(n-1)^2}{(n-2)(n-3)}$$

で定義されています。ここで s は標準偏差で、データ数 n が 3 以下の時エラーとなります。

R では尖度を求める関数が定義されていないので、自分で定義することになります。library“e1071"を組み込むと、歪度の計算に関して `kurtosis` という関数を利用することが出来ます[*4]。

関数 `kurt` の定義

```
> kurt<- function(x){                    # 関数 kurt の定義。
>   m4<- sum((x-mean(x))^4)/length(x)
>   s4<- var(x)^2
>   m4/s4-3
> }
> kurt(x)                                # kurt の値の出力。
```

3.4.3　変動係数

変動係数(coefficient of variation) は、データ $X_1, X_2, \cdots, X_n$ の平均 $\overline{X}$ と標準偏差 s_X を用いて

$$CV=\frac{s_X}{\overline{X}}$$

によって定義され、データの単位に無関係な量を与えている。つまり、X_i の単位が cm のとき、平均 $\overline{X}$ と標準偏差 s_X の単位はそれぞれ、cm で変動係数 CV の単位は無単位となる。

R では変動係数を求める関数が定義されていないので、自分で定義することになります。

関数 `CV` の定義

```
> CV<- function(x) sd(x)/mean(x)    # 関数 CV の定義。
> CV(x)                             # CV の値の出力。
```

3.4.4　Gini の係数

Gini の係数(Gini coefficient) は、

$$GI=\frac{1}{2n^2\overline{X}}\sum_{i=1}^{n}\sum_{j=1}^{n}|X_i-X_j|$$

で与えられるもので、不平等度や集中度の指標として用いられる。

[*4] ライブラリについては付録 C のライブラリのインストール参照。

関数 GI の定義

```
> GI<- function(x)                    # 関数 GI の定義。
> {
>  n<- length(x)
>  y<- sort(x)
>  j<- seq(1,n)
>  2*sum(j*y)/(n^2*mean(x))-(n+1)/n
> }
> GI(x)                               # GI の値の出力。
```

R では変動係数を求める関数が定義されていないので、上の式を変形して、定義することになります。

問題

問題 3.1.

データ iris は 3 種類のアヤメ (setosa、versicolor、virginica) について、ガク片の長さ・幅、花片の長さ・幅を測定した、各種について 50 個、計 150 個のデータである。setosa と versicolor のガク片の長さについて、ヒストグラムを描き、算術平均、幾何平均、調和平均、中央値、最大値、最小値、下側ヒンジ、上側ヒンジ、第 1 四分位点、第 3 四分位点、分散、標準偏差、範囲、ヒンジ散布度、四分位範囲を求めよ。また、setosa のガク片の長さと幅の相関係数と共分散を求めよ。

入力のヒント

```
>  data(iris)
> x<- iris[iris$Species=="setosa",1]      # データ iris の種 setosa のガク片の長さを x とする。
> y<- iris[iris$Species=="versicolor",1]  # データ iris の種 versicolor のガク片の長さを y とする。
```

問題 3.2.

関数 fivenum と summary は seq(1,6) に対して異なる結果を導くことを確認せよ。

問題 3.3.

正規乱数 rnorm を用いて、歪度についての下記のコマンドを実行し、コマンドの意味の解釈を与えよ。

入力

```
> sym <- rnorm(100)
> neg <- sym[sym<0]
> pos <- sym[sym>0]
> neg.skew <- c(-(neg^2), pos^.5)
> pos.skew <- c(-((-neg)^.5), pos^2)
> s.levels <- rep(c("neg.skew", "sym", "pos.skew"), each=100)
> d <- data.frame(dist=s.levels, y=c(neg.skew, sym, pos.skew))
> boxplot(y~dist, data=d)
```

Tips 履歴について

入力してきた履歴を見るのには history() というコマンドを用います。

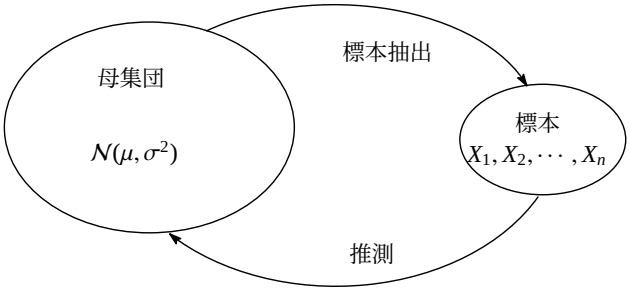
入力

```
>history()    # 全履歴の表示。
```

また、キーボートの上向きのキー↑を押すことで直前までに入力したコマンドを順に表示することが出来、Enter キー↵を押すことでコマンドが実行されます。

第 4 章　一標本正規母集団に関する検定

母集団が**正規分布**(normal distribution) であると仮定できる状況を考える。正規分布は平均 μ と分散 σ^2 から一意に決定される分布で、この二つの**母数**(パラメータ：parameter) についての幾つかの検定問題を考える。



母集団から大きさ n の標本 $X_1, X_2, \cdots, X_n$ が抽出され、この標本に基づき母集団の母数について検定する。

検定の手順

1. 分布、モデル等の前提条件のチェック。
2. 帰無仮説と対立仮説、有意水準 (α) の設定。
3. 棄却領域の設定。
4. 検定統計量の計算。
5. 帰無仮説を棄却するまたは棄却しないの判定と結論。

母数の値について既知であるという場合と未知であるという場合がある。値を**既知**(known) と仮定できるのは今までの経験から値が分かっている状況で、**未知**(unknown) と仮定するのは今までと違う性質のものを扱っているという状況において起こるものである。

4.1　平均の検定

母分散 σ^2 の正規母集団から標本 $X_1, X_2, \cdots, X_n$ が抽出されている。有意水準は α とし、以下の三つの検定問題を扱うことになる。

両側検定 (two-sided test)　対立仮説 $\mathcal{H}_1$: $\mu \neq \mu_0$ に対して帰無仮説 $\mathcal{H}_0$: $\mu = \mu_0$ を検定する問題

片側検定 (one-sided test)
- 対立仮説 $\mathcal{H}_1$: $\mu > \mu_0$ に対して帰無仮説 $\mathcal{H}_0$: $\mu = \mu_0$ を検定する問題
- 対立仮説 $\mathcal{H}_1$: $\mu < \mu_0$ に対して帰無仮説 $\mathcal{H}_0$: $\mu = \mu_0$ を検定する問題

ここで、母分散 σ^2 が既知の場合と未知の場合について分けて扱う必要がある。

4.1.1　母分散 σ^2 が既知の場合

ここでの検定問題について共通な事柄は、標本 $X_1, X_2, \cdots, X_n$ に対する標本平均

$$\overline{X} = \frac{1}{n}\{X_1 + X_2 + \cdots + X_n\} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

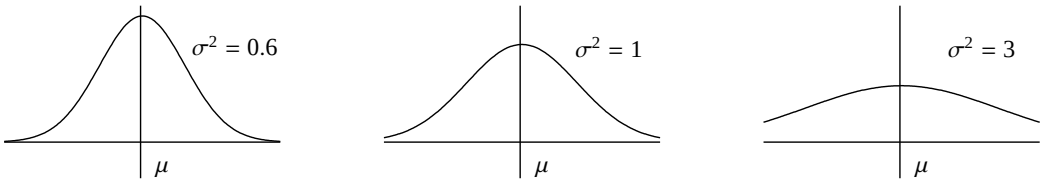
を求め、標準化した値 (検定統計量)

$$\begin{aligned} Z &= \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} \\ &= \frac{\sum_{i=1}^{n} X_i/n - \mu_0}{\sigma/\sqrt{n}} \end{aligned} \tag{4.1}$$

を計算するということになる。帰無仮説 $\mathcal{H}_0$: $\mu = \mu_0$ の下で抽出された標本は正規分布 $\mathcal{N}(\mu_0, \sigma^2)$ に従うので、検定統計量 Z は標準正規分布 $\mathcal{N}(0, 1^2)$ に従うことが知られている。ここで、平均 μ、分散 σ^2 の正規分布の分布関数は

$$\int_{-\infty}^{x} \frac{1}{\sqrt{2\sigma^2}} \exp(-\frac{(z-\mu)^2}{2})\, dz$$

で与えられており、被積分関数である正規密度関数は平均 μ と分散 σ^2 の二つの量によって一意に決まっている。[*1]



検定統計量 Z の値により帰無仮説からの離脱が分かるので、下記のような棄却領域が作られる。

帰無仮説 $\mathcal{H}_0$: $\mu = \mu_0$ *versus* 対立仮説 $\mathcal{H}_1$: $\mu \neq \mu_0$
$$\begin{cases} |Z| > z_{\alpha/2} \text{ ならば帰無仮説を棄却する。} \\ |Z| \leq z_{\alpha/2} \text{ ならば帰無仮説は棄却できない。} \end{cases}$$

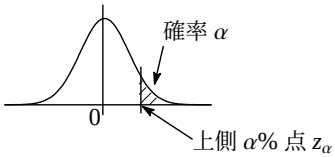
帰無仮説 $\mathcal{H}_0$: $\mu = \mu_0$ *versus* 対立仮説 $\mathcal{H}_1$: $\mu > \mu_0$
$$\begin{cases} Z > z_{\alpha} \text{ ならば帰無仮説を棄却する。} \\ Z \leq z_{\alpha} \text{ ならば帰無仮説は棄却できない。} \end{cases}$$

帰無仮説 $\mathcal{H}_0$: $\mu = \mu_0$ *versus* 対立仮説 $\mathcal{H}_1$: $\mu < \mu_0$
$$\begin{cases} Z < -z_{\alpha} \text{ ならば帰無仮説を棄却する。} \\ Z \geq -z_{\alpha} \text{ ならば帰無仮説は棄却できない。} \end{cases}$$

ここで、z_α は標準正規分布 $\mathcal{N}(0, 1^2)$ の上側 α% 点で

$$\alpha = \int_{z_\alpha}^{\infty} \frac{1}{\sqrt{2\pi}} \exp(-\frac{x^2}{2})\, dx$$

である。



例題 4.1. R のデータ `iris` の種 `setosa` のガク片の長さを考え、ガク片の長さの平均についての仮説検定を扱う。過去の経験から母分散は 0.11(標準偏差は 0.3316625) であることが分かっていると仮定する。有意水準は 0.05 とする。ヒストグラムと描き、このデータから計算された平均と分散を持つ正規分布を重ね合わせたものが図 4.1 で、正規分布に従っていると仮定しても差し支えなさそうである[*2]。■

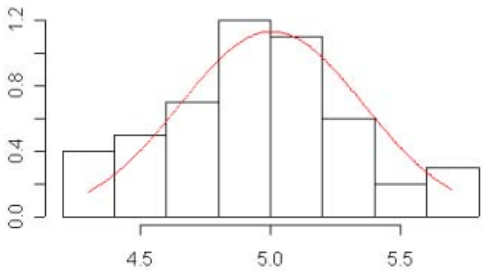


図 4.1　アヤメの種 setosa のガク片の長さのヒストグラム

[*1] 例えば、`mu<- 1; s<- 10; x<- seq(mu-3*s, mu+3*s, by=0.1);plot(x, dnorm(x,mu,s), type="l")` により、平均 1、分散 10^2 の正規分布密度関数を描くことが出来る。

[*2] 第 6 章で扱う適合度検定等で正規性を検定する必要がある。

(I) 両側検定: 帰無仮説 $\mathcal{H}_0$: $\mu = 5.5$ *versus* 対立仮説 $\mathcal{H}_1$: $\mu \neq 5.5$ (分散 σ^2 は既知で 0.11 と仮定)
統計量

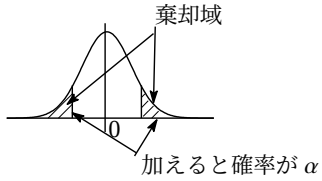
$$Z = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}}$$

の絶対値が棄却限界値 $z_{0.025} = 1.96$ より大きいとき有意水準 5% で帰無仮説を棄却する。

入力と出力結果

```
> data(iris)
> x<- iris[iris$Species=="setosa",1]   # データ iris の種 setosa のガク片の長さを x とする。
> mean(x)                              # x の平均の計算。
[1] 5.006
> (mean(x)-5.5)/sqrt(0.11/50)          # データ数 50、平均 5.5、分散 0.11 としたときの値。
[1] -10.53212
```

setosa のガク片の長さの平均は 5.006 であることが分かり、標準化した値は −10.53212、絶対値を取れば 10.53212 となる。棄却限界値 $z_{0.025} = 1.96$ と比べて大きいので、有意水準 0.05 で帰無仮説が棄却される (R の関数 qnorm(1-0.025) を用いると 1.959964 という値が求まり、これは $z_{0.025}$ の有効桁数 7 桁のときの値である)。



(II) 片側検定: 帰無仮説 $\mathcal{H}_0$: $\mu = 4.8$ *versus* 対立仮説 $\mathcal{H}_1$: $\mu > 4.8$ (分散 σ^2 は既知で 0.11 と仮定)
標本 $X_1, X_2, \cdots, X_n$ から計算される検定統計量

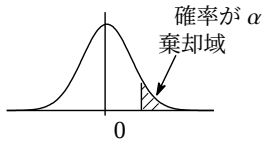
$$Z = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}}$$

の値が棄却限界値 $z_{0.05} = 1.64$ より大きいとき有意水準 5% で帰無仮説を棄却する。

入力と出力結果

```
> data(iris)
> x<- iris[iris$Species=="setosa",1]   # データ iris の種 setosa のガク片の長さを x とする。
> mean(x)                              # x の平均の計算。
[1] 5.006
> (mean(x)-4.8)/sqrt(0.11/50)          # データ数 50、平均 4.8、分散 0.11 とした検定統計量の値。
[1] 4.391935
```

Z の値は 4.391935 で、これは棄却限界値 $z_{0.05} = 1.64$(R では qnorm(1-0.05)=1.644854) に比べ大きいので、有意水準 0.05 で帰無仮説は棄却される。



(III) 片側検定: 帰無仮説 $\mathcal{H}_0$: $\mu = 5.0$ *versus* 対立仮説 $\mathcal{H}_1$: $\mu < 5.0$ (分散 σ^2 は既知で 0.11 と仮定)
標本 $X_1, X_2, \cdots, X_n$ から計算される検定統計量

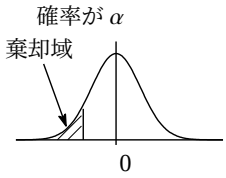
$$Z = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}}$$

の値が棄却限界値 $z_{0.95} = -1.64$ より小さいとき有意水準 5% で帰無仮説を棄却する。

入力と出力結果

```
> data(iris)
> x<- iris[iris$Species=="setosa",1]   # データ iris の種 setosa のガク片の長さを x とする。
> mean(x)                              # x の平均の計算。
[1] 5.006
> (mean(x)-5)/sqrt(0.11/50)            # データ数 50、平均 5、分散 0.11 としたときの値。
[1] 0.1279204
```

Z の値は 0.1279204 で、これは棄却限界値 -1.64(R では `qnorm(0.05)`$= -1.644854$) に比べ大きいので、有意水準 0.05 で帰無仮説は棄却されない。



4.1.2　母分散 σ^2 が未知の場合

未知の母分散 σ^2 を持つ正規母集団から標本 $X_1, X_2, \cdots, X_n$ が抽出されている。有意水準は α とし、以下の三つの検定問題を扱う。

両側検定　対立仮説 $\mathcal{H}_1$: $\mu \neq \mu_0$ に対して帰無仮説 $\mathcal{H}_0$: $\mu = \mu_0$ を検定する問題

片側検定 $\begin{cases} \text{対立仮説 } \mathcal{H}_1\text{: } \mu > \mu_0 \text{ に対して帰無仮説 } \mathcal{H}_0\text{: } \mu = \mu_0 \text{ を検定する問題} \\ \text{対立仮説 } \mathcal{H}_1\text{: } \mu < \mu_0 \text{ に対して帰無仮説 } \mathcal{H}_0\text{: } \mu = \mu_0 \text{ を検定する問題} \end{cases}$

母分散 σ^2 が未知であるので、前節の手法を用いることは出来ないので、式 (4.1) の Z の計算式に現れる未知の量 σ^2 を標本分散（不偏分散）

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

で置き換えた統計量

$$t = \frac{\overline{X} - \mu_0}{\frac{s}{\sqrt{n}}}$$

を用いる[*3]。標本が正規分布 $\mathcal{N}(\mu_0, \sigma^2)$ から抽出されたとき、統計量 t は自由度 $n-1$ の t 分布に従うということが知られている。ここで、自由度 n の t 分布関数は

$$\int_{-\infty}^{x} \frac{\Gamma((n+1)/2)}{\Gamma(1/2)\Gamma(n/2)}\left(1 + \frac{t^2}{n}\right)^{-(n+1)/2} dt$$

である。この統計量 t を用いると、下記のような棄却領域が作られる。

帰無仮説 $\mathcal{H}_0$: $\mu = \mu_0$ *versus* 対立仮説 $\mathcal{H}_1$: $\mu \neq \mu_0$ $\begin{cases} |t| > t_{n-1}(\alpha/2) \text{ ならば帰無仮説を棄却する。} \\ |t| \leq t_{n-1}(\alpha/2) \text{ ならば帰無仮説は棄却できない。} \end{cases}$

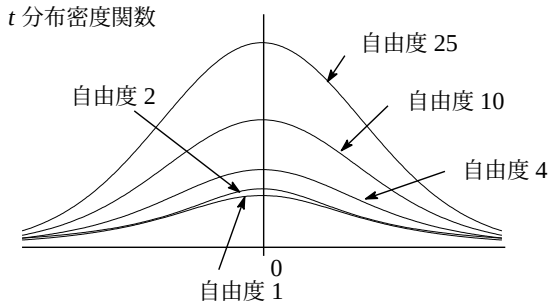
帰無仮説 $\mathcal{H}_0$: $\mu = \mu_0$ *versus* 対立仮説 $\mathcal{H}_1$: $\mu > \mu_0$ $\begin{cases} t > t_{n-1}(\alpha) \text{ ならば帰無仮説を棄却する。} \\ t \leq t_{n-1}(\alpha) \text{ ならば帰無仮説は棄却できない。} \end{cases}$

帰無仮説 $\mathcal{H}_0$: $\mu = \mu_0$ *versus* 対立仮説 $\mathcal{H}_1$: $\mu < \mu_0$ $\begin{cases} t < -t_{n-1}(\alpha) \text{ ならば帰無仮説を棄却する。} \\ t \geq -t_{n-1}(\alpha) \text{ ならば帰無仮説は棄却できない。} \end{cases}$

ここで、$t_n(\alpha)$ は自由度 n の t 分布の上側 α% 点で

$$\alpha = \int_{t_n(\alpha)}^{\infty} \frac{\Gamma((n+1)/2)}{\Gamma(1/2)\Gamma(n/2)}\left(1 + \frac{x^2}{n}\right)^{-(n+1)/2} dx$$

である。この被積分関数（自由度 n の t 分布密度関数）は、自由度 n とともに変化し、



自由度 1,2,4,10,25 の t 分布密度関数

```
x<- seq(-4,4,by=0.2)
plot(c(-4,4),c(0,0.5),type="n")
j<- 0
for (i in c(1,2,4,10,25))
{  j<- j+1
   lines(spline(x,dt(x,i)),col=j)}
```

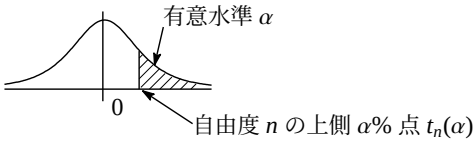
[*3] 不偏分散については 25 ページ参照。

という形をしている。

上の式で用いられる関数 Γ は、ガンマ関数と呼ばれ、

$$\Gamma(n) = \int_0^\infty x^{n-1}\exp(-x)\,dx$$

によって与えられ、$\Gamma(\frac{1}{2}) = \sqrt{\pi}, \Gamma(1) = 1, \Gamma(n) = (n-1)!$ という性質をもつ[*4]。



例題 4.2. R のデータ `iris` の種 `setosa` のガク片の長さを考え、ガク片の長さの平均についての仮説検定を扱う。有意水準は 0.05 とする。ただし、分散は未知であるとする。■

(I) 両側検定: 帰無仮説 $\mathcal{H}_0$: $\mu = 5.5$ *versus* 対立仮説 $\mathcal{H}_1$: $\mu \neq 5.5$ (分散は未知)

$$\text{検定手法:} \qquad |t| = \left|\frac{\overline{X} - \mu_0}{s/\sqrt{n}}\right| > t_{n-1}(\alpha/2) \Rightarrow \mathcal{H}_0 : \mu = \mu_0 \text{ を棄却する。}$$

入力と出力結果

```
> data(iris)
> x<- iris[iris$Species=="setosa",1]                  # データ iris の種 setosa のガク片の長さを x とする。
> t.test(x, mu=5.5, alternative="two.sided")          # t-検定。5.5 は帰無仮説の値。対立仮説は両側。
        One Sample t-test                             # 一標本検定
data:  x                                              # データ名：x
t = -9.9098, df = 49, p-value = 2.717e-13             # t の値、自由度=49、p-値。
alternative hypothesis: true mean is not equal to 5   # 対立仮説：H1 : μ ≠ 5。
95 percent confidence interval:                       # 95% 信頼区間
 4.905824 5.106176                                    #   [4.905824, 5.106176]
sample estimates:                                     # 標本推定値。
mean of x                                             # データ x の平均
    5.006                                             #   5.006
> (mean(x)-5.5)/sqrt(var(x)/length(x))                # 関数 t.test を用いない方法。
[1] -9.909815
```

- t の値は (mean(x)-5.5)/sqrt(var(x)/length(x)) の値と一致している。
- 自由度は標本数から 1 を引いたものである。
- p-値 (p-value) とは、考えている対立仮説に対して、帰無仮説の下で計算された t 値が起こる確率を表す。よって、この値が有意水準より小さな値になるとき帰無仮説が棄却されることになり、今の場合、有意水準 0.05 で帰無仮説は棄却される。通常は t 分布表から求まる $t_{49}(0.025)$ の値と t の値を比較するが R では p-値が出力されるので、出力結果に $t_{49}(0.025)$ の値は検定において現れない。
- 95% 信頼区間とは上で用いた標本から、母集団平均が区間 [4.905824, 5.106176] に含まれる確率が 95% であるというものである。
- 母平均の点推定値は 5.006 である。

(II) 片側検定: 帰無仮説 $\mathcal{H}_0$: $\mu = 4.8$ *versus* 対立仮説 $\mathcal{H}_1$: $\mu > 4.8$ (分散は未知)

$$\text{検定手法:} \qquad t = \frac{\overline{X} - \mu_0}{s/\sqrt{n}} > t_{n-1}(\alpha) \Rightarrow \mathcal{H}_0 : \mu = \mu_0 \text{ を棄却する。}$$

[*4] 例えば、`n<- 8; x<- seq(-5, 5, by=0.1);plot(x, dt(x,df=n), type="l")` により、自由度 n の t 分布密度関数を描くことが出来る。

入力と出力結果

```
> data(iris)
> x<- iris[iris$Species=="setosa",1]                    # データ iris の種 setosa のガク片の長さを x とする。
> t.test(x, mu=4.8, alternative="greater")              # t-検定。4.8 は帰無仮説の値。対立仮説は片側。
        One Sample t-test                                # 一標本検定
data:  x                                                 # データ名：x
t = 4.1324, df = 49, p-value = 6.986e-05                 # t の値、自由度=49、p-値。
alternative hypothesis: true mean is greater than 4.8    # 対立仮説：H1 : μ > 4.8。
95 percent confidence interval:                          # 95% 信頼区間。
 4.922425      Inf                                       # [4.922425, +∞)
sample estimates:                                        # 標本推定値。
mean of x                                                # データ x の平均。
    5.006                                                #   5.006
> (mean(x)-4.8)/sqrt(var(x)/length(x))                   # 関数 t.test を用いない方法。
[1] 4.132433
```

- p-値はほぼゼロであるので、帰無仮説の下で標本が抽出されているとは考えられなく、p-値が 0.05 より小さいので、有意水準 0.05 で帰無仮説は棄却される。
- 95% 信頼区間とは上で用いた標本から、母集団平均が区間 $[4.922425, +\infty)$ に含まれる確率が 95% であるというものである。

(III) 片側検定: 帰無仮説 $\mathcal{H}_0$: $\mu = 5.0$ *versus* 対立仮説 $\mathcal{H}_1$: $\mu < 5.0$ (分散は未知)

$$\text{検定手法:}\qquad t = \frac{\overline{X} - \mu_0}{s/\sqrt{n}} < -t_{n-1}(\alpha) \Rightarrow \mathcal{H}_0 : \mu = \mu_0 \text{ を棄却する。}$$

入力と出力結果

```
> data(iris)
> x<- iris[iris$Species=="setosa",1]                    # データ iris の種 setosa のガク片の長さを x とする。
 > t.test(x, mu=4.8, alternative="less")                # t-検定。4.8 は帰無仮説の値。対立仮説は片側。
        One Sample t-test                                # 一標本検定
data:  x                                                 # データ名：x
t = 0.1204, df = 49, p-value = 0.5477                    # t の値、自由度=49、p-値。
alternative hypothesis: true mean is less than 5         # 対立仮説：H1 : μ < 5
95 percent confidence interval:                          # 95% 信頼区間
     -Inf 5.089575                                       # (−∞, 5.089575)
sample estimates:                                        # 標本推定値。
mean of x                                                # データ x の平均。
    5.006                                                #   5.006
> (mean(x)-4.8)/sqrt(var(x)/length(x))                   # 関数 t.test を用いない方法。
[1] 0.1203621
```

- p-値が有意水準より大きな値であるので、有意水準 0.05 では帰無仮説を棄却できない。

4.2　母分散 σ^2 の検定

未知の母分散 σ^2 を持つ正規母集団から標本 $X_1, X_2, \cdots, X_n$ が抽出されている。有意水準は α とし、以下の三つの検定問題を扱う。

$$\text{両側検定}\quad \text{対立仮説 } \mathcal{H}_1\text{: } \sigma^2 \neq \sigma_0^2 \text{ に対して帰無仮説 } \mathcal{H}_0\text{: } \sigma^2 = \sigma_0^2 \text{ を検定する問題}$$

$$\text{片側検定}\begin{cases} \text{対立仮説 } \mathcal{H}_1\text{: } \sigma^2 > \sigma_0^2 \text{ に対して帰無仮説 } \mathcal{H}_0\text{: } \sigma^2 = \sigma_0^2 \text{ を検定する問題} \\ \text{対立仮説 } \mathcal{H}_1\text{: } \sigma^2 < \sigma_0^2 \text{ に対して帰無仮説 } \mathcal{H}_0\text{: } \sigma^2 = \sigma_0^2 \text{ を検定する問題} \end{cases}$$

4.2.1　平均 μ が既知の場合

母平均 μ が既知、母分散 σ^2 が未知であるので、標本 $X_1, X_2, \cdots, X_n$ から計算される分散

$$\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2$$

を用いて母分散 σ^2 を推定する。分母は $n-1$ でなく、n を用いる。この分散は標本数が大きくなると母分散 σ^2 の値に近づくので、帰無仮説 $\mathcal{H}_0$: $\sigma^2=\sigma_0^2$ を検定するには

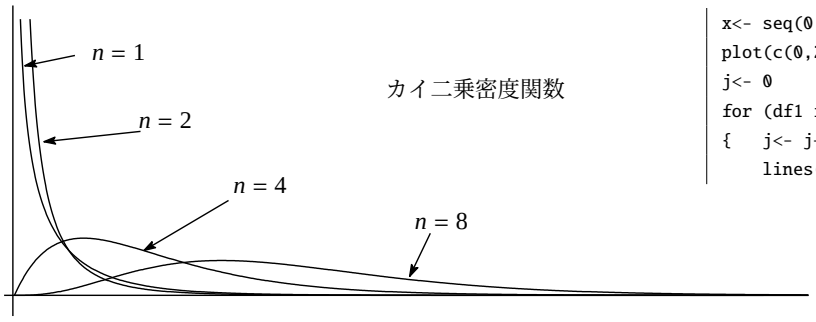
$$\chi_0^2=\frac{\sum_{i=1}^{n}(X_i-\mu)^2}{\sigma_0^2}$$

を検定統計量として用いることになる。

帰無仮説の下で、この χ_0^2 は自由度 n のカイ二乗分布に従うことが知られている。ここで、自由度 n のカイ二乗分布関数は

$$\int_0^x \frac{1}{2^{n/2}\Gamma(n/2)}t^{n/2-1}\exp(-\frac{t}{2})\,dt$$

で与えられ、被積分関数は、自由度 n とともに変化し、



自由度 1, 2, 4, 8 のカイ二乗密度関数

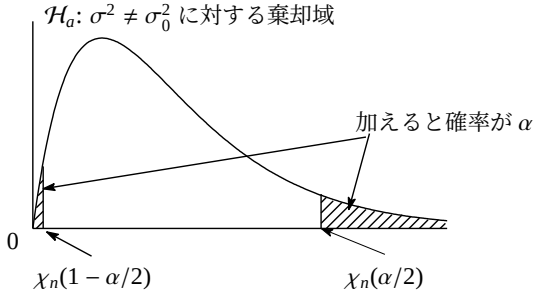
```
x<- seq(0.1,25,0.2)
plot(c(0,25),c(0,1.5),type="n")
j<- 0
for (df1 in c(1,2,4,8))
{   j<- j+1
    lines(spline(x,dchisq(x,df1)),col=j)}
```

という形をしている。

これにより、有意水準 α に対して下記のような棄却領域が作られる。

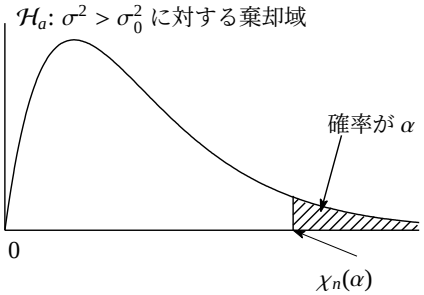
(I) 帰無仮説 $\mathcal{H}_0$: $\sigma^2=\sigma_0^2$ *versus* 対立仮説 $\mathcal{H}_1$: $\sigma^2\neq\sigma_0^2$

検定統計量 χ_0^2 の値が、小さすぎるときまたは、大きすぎるときに帰無仮説が起きているとはいえないので、$\chi_0^2<\chi_n(1-\alpha/2)$ または $\chi_0^2>\chi_n(\alpha/2)$ ならば帰無仮説を棄却する。



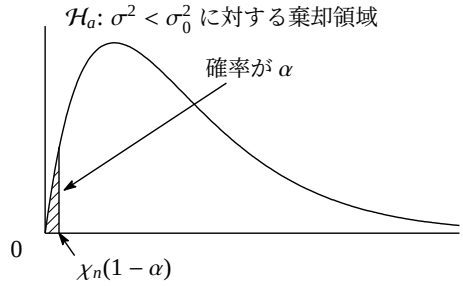
(II) 帰無仮説 $\mathcal{H}_0$: $\sigma^2=\sigma_0^2$ *versus* 対立仮説 $\mathcal{H}_1$: $\sigma^2>\sigma_0^2$

検定統計量 χ_0^2 の値が、大きすぎるとき帰無仮説が起きているとはいえないので、$\chi_0^2>\chi_n(\alpha)$ ならば帰無仮説を棄却する。



(III) 帰無仮説 $\mathcal{H}_0$: $\sigma^2 = \sigma_0^2$ *versus* 対立仮説 $\mathcal{H}_1$: $\sigma^2 < \sigma_0^2$

検定統計量 χ_0^2 の値が、小さすぎるとき帰無仮説が起きているとはいえないので、$\chi_0^2 < \chi_n(1-\alpha)$ ならば帰無仮説を棄却する。



ここで、$\chi_n(\alpha)$ は自由度 n のカイ二乗分布 χ_n^2 の上側 α% 点で

$$\alpha = \int_{\chi_n(\alpha)}^{+\infty} \frac{1}{2^{n/2}\Gamma(\frac{n}{2})} t^{n/2-1} \exp(-\frac{t}{2})\, dt$$

である。

例題 4.3. R のデータ `iris` の種 `setosa` のガク片の長さを考え、ガク片の長さの分散についての仮説検定を扱う。有意水準は 5% とし、平均は 5 だと分かっているものとする。■

(I) 両側検定: 帰無仮説 $\mathcal{H}_0$: $\sigma^2 = 0.2$ *versus* 対立仮説 $\mathcal{H}_1$: $\sigma^2 \neq 0.2$ (平均は 5 であると分かっている)

検定手法:　$\chi_0^2 < \chi_n(1-\alpha/2)$ または$\chi_0^2 > \chi_n(\alpha/2) \Rightarrow \mathcal{H}_0 : \sigma^2 = \sigma_0^2$ を棄却する。

入力と出力結果

```
> data(iris)
> x<- iris[iris$Species=="setosa",1]    # データ iris の種 setosa のガク片の長さを x とする。
> sum((x-5)^2)/0.2                      # χ0^2 の値。
[1] 30.45
```

- 検定統計量 χ_0^2 の値は 30.45 で、自由度 50、有意水準 0.05 に対して下側棄却限界値は $\chi_{50}(1-0.05/2) = 32.35736$ で上側棄却限界値 $\chi_{50}(0.05/2) = 71.4202$ である。ここで、下側棄却限界値は `qchisq(0.025,df=50)=32.35736` から求まり、上側棄却限界値は `qchisq(0.975,df=50)=71.4202` から求まる。よって、χ_0^2 の値 30.45 は下側棄却限界値 32.35736 より小さいので帰無仮説 $\mathcal{H}_0$: $\sigma^2 = 0.2$ は有意水準 0.05 で棄却される。

(II) 片側検定: 帰無仮説 $\mathcal{H}_0$: $\sigma^2 = 0.15$ *versus* 対立仮説 $\mathcal{H}_1$: $\sigma^2 > 0.15$ (平均は 5 であると分かっている)

検定手法:　$\chi_0^2 > \chi_n(\alpha) \Rightarrow \mathcal{H}_0 : \sigma^2 = \sigma_0^2$ を棄却する。

入力と出力結果

```
> data(iris)
> x<- iris[iris$Species=="setosa",1]    # データ iris の種 setosa のガク片の長さを x とする。
> sum((x-5)^2)/0.15                     # χ0^2 の値。
[1] 40.55
```

- 検定統計量 χ_0^2 の値は 40.55 で、自由度 50、有意水準 0.05 に対して棄却限界値は $\chi_{50}(0.05) = 67.5048$ である。ここで、棄却限界値は `qchisq(0.95,df=50)=67.5048` から求まる。よって、χ_0^2 の値 40.55 は棄却限界値 67.5048 より小さいので帰無仮説 $\mathcal{H}_0$: $\sigma^2 = 0.15$ は有意水準 0.05 で棄却されない。

(III) 片側検定: 帰無仮説 $\mathcal{H}_0$: $\sigma^2 = 0.1$ *versus* 対立仮説 $\mathcal{H}_1$: $\sigma^2 < 0.1$ (平均は 5 であると分かっている)

検定手法:　$\chi_0^2 < \chi_n(1-\alpha) \Rightarrow \mathcal{H}_0 : \sigma^2 = \sigma_0^2$ を棄却する。

入力と出力結果

```
> data(iris)
> x<- iris[iris$Species=="setosa",1]    # データ iris の種 setosa のガク片の長さを x とする。
> sum((x-5)^2)/0.1                      # χ0^2 の値。
[1] 60.9
```

- 検定統計量 χ_0^2 の値は 60.9 で、自由度 50、有意水準 0.05 に対して下側棄却限界値は $\chi_{50}(0.95) = 34.76425$ である。ここで、下側棄却限界値は `qchisq(0.05,df=50)=34.76425` から求まる。よって、χ_0^2 の値 60.9 は下側棄却限界値 34.76425 より大きいので、帰無仮説 $\mathcal{H}_0$: $\sigma^2 = 0.1$ は有意水準 0.05 で棄却されない。

4.2.2　平均 μ が未知の場合

母平均 μ が未知であるので、前節において利用した

$$\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2$$

の代わりに

$$\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

を用いて母分散 σ^2 を推定する。母集団平均 μ でなく、標本平均 $\overline{X}$ を用いるので分母は n でなく、$n-1$ になる。この分散は標本数が大きくなると母分散 σ^2 の値に近づくので、帰無仮説 $\mathcal{H}_0$: $\sigma^2 = \sigma_0^2$ を検定するには

$$\chi_0^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{\sigma_0^2}$$

を検定統計量として用いる。この値が 1 に近いと帰無仮説からの標本と考えてよい。この検定統計量 χ_0^2 は自由度 $n-1$ のカイ二乗分布に従うという性質を持っている。前節では平均が分かっていたので

$$\frac{\sum_{i=1}^{n}(X_i - \mu)^2}{\sigma_0^2}$$

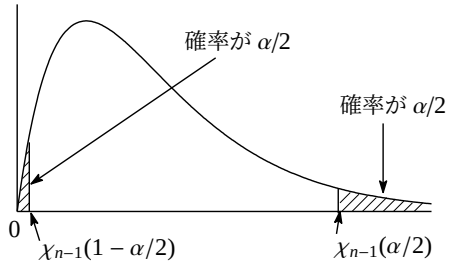
を用いたが、未知の μ の代わりに $\overline{X}$ を用いることによりより、自由度が 1 減っている。

以上のことから、有意水準 α に対して下記のような棄却領域が作られる。

(I) 帰無仮説 $\mathcal{H}_0$: $\sigma^2 = \sigma_0^2$ *versus* 対立仮説 $\mathcal{H}_1$: $\sigma^2 \neq \sigma_0^2$

検定統計量 χ_0^2 の値が、小さすぎるときまたは、大きすぎるときに帰無仮説が起きているとはいえないので、$\chi_0^2 < \chi_{n-1}(1-\alpha/2)$ または $\chi_0^2 > \chi_{n-1}(\alpha/2)$ ならば帰無仮説を棄却する。

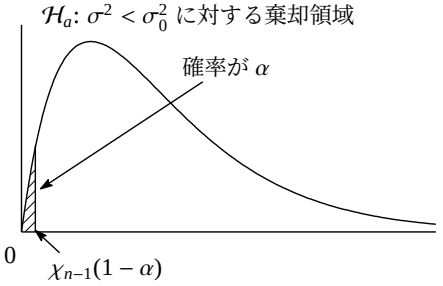
$\mathcal{H}_a$: $\sigma^2 \neq \sigma_0^2$ に対する棄却域



(II) 帰無仮説 $\mathcal{H}_0$: $\sigma^2 = \sigma_0^2$ *versus* 対立仮説 $\mathcal{H}_1$: $\sigma^2 > \sigma_0^2$

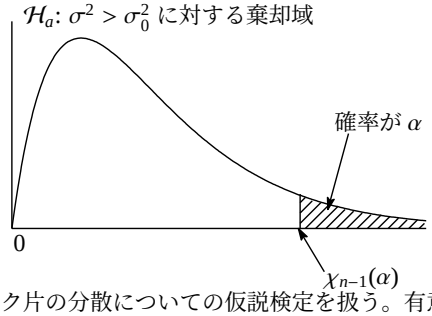
検定統計量 χ_0^2 の値が、大きすぎるときに帰無仮説が起きているとはいえないので、$\chi_0^2 > \chi_{n-1}(\alpha)$ ならば帰無仮説を棄却する。

$\mathcal{H}_a$: $\sigma^2 < \sigma_0^2$ に対する棄却領域



(III) 帰無仮説 $\mathcal{H}_0$: $\sigma^2 = \sigma_0^2$ *versus* 対立仮説 $\mathcal{H}_1$: $\sigma^2 < \sigma_0^2$

検定統計量 χ_0^2 の値が、小さすぎるときに帰無仮説が起きているとはいえないので、$\chi_0^2 < \chi_{n-1}(1-\alpha)$ ならば帰無仮説を棄却する。



例題 4.4. R のデータ `iris` の種 `setosa` のガク片の長さを考え、ガク片の分散についての仮説検定を扱う。有意水準は 5% とし、平均は未知であるとする。■

(I) 両側検定: 帰無仮説 $\mathcal{H}_0$: $\sigma^2 = 0.2$ *versus* 対立仮説 $\mathcal{H}_1$: $\sigma^2 \neq 0.2$ (平均は未知)

検定手法:　$\chi_0^2 < \chi_n(1-\alpha/2)$ または$\chi_0^2 > \chi_n(\alpha/2) \Rightarrow \mathcal{H}_0 : \sigma^2 = \sigma_0^2$ を棄却する。

入力と出力結果

```
> data(iris)
> x<- iris[iris$Species=="setosa",1]    # データ iris の種 setosa のガク片の長さを x とする。
> sum((x-mean(x))^2)/0.2                # χ0^2 の値。
[1] 30.441
```

- 検定統計量 χ_0^2 の値は 30.441 で、自由度 49、有意水準 0.05 に対して、下側棄却限界値は $\chi_{49}(1-0.05/2) = 31.55492$ で上側棄却限界値は $\chi_{49}(0.05/2) = 70.22241$ である。ここで、下側棄却限界値は `qchisq(0.025,df=49)=31.55492` から求まり、上側棄却限界値は `qchisq(0.975,df=49)=70.22241` から求まる。よって、χ_0^2 の値 30.441 は下側棄却限界値 31.55492 より小さいので帰無仮説 $\mathcal{H}_0$: $\sigma^2 = 0.2$ は有意水準 0.05 で棄却される。

(II) 片側検定: 帰無仮説 $\mathcal{H}_0$: $\sigma^2 = 0.15$ *versus* 対立仮説 $\mathcal{H}_1$: $\sigma^2 > 0.15$ (平均は未知)

検定手法:　$\chi_0^2 > \chi_n(\alpha) \Rightarrow \mathcal{H}_0 : \sigma^2 = \sigma_0^2$ を棄却する。

入力と出力結果

```
> data(iris)
> x<- iris[iris$Species=="setosa",1]    # データ iris の種 setosa のガク片の長さを x とする。
> sum((x-mean(x))^2)/0.15               # χ0^2 の値。
[1] 40.588
```

- 検定統計量 χ_0^2 の値は 40.588 で、自由度 49、有意水準 0.05 に対して、棄却限界値は $\chi_{49}(0.05) = 66.33865$ である。ここで、棄却限界値は `qchisq(0.95,df=49)=66.33865` から求まる。よって、χ_0^2 の値 40.588 は棄却限界値 66.33865 より小さいので帰無仮説 $\mathcal{H}_0$: $\sigma^2 = 0.2$ は有意水準 0.05 で棄却されない。

(III) 片側検定: 帰無仮説 $\mathcal{H}_0$: $\sigma^2 = 0.1$ *versus* 対立仮説 $\mathcal{H}_1$: $\sigma^2 < 0.1$ (平均は未知)

検定手法:　$\chi_0^2 < \chi_n(1-\alpha) \Rightarrow \mathcal{H}_0 : \sigma^2 = \sigma_0^2$ を棄却する。

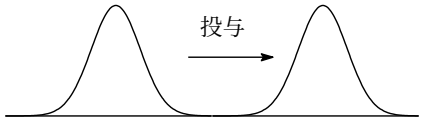
入力と出力結果

```
> data(iris)
> x<- iris[iris$Species=="setosa",1]    # データ iris の種 setosa のガク片の長さを x とする。
> sum((x-mean(x))^2)/0.1                # χ0^2 の値。
[1] 60.882
```

- 検定統計量 χ_0^2 の値は 60.882 で、自由度 49、有意水準 0.05 に対して、棄却限界値は $\chi_{49}(0.95) = 33.93031$ である。ここで、棄却限界値は `qchisq(0.05,df=49)=33.93031` から求まる。よって、χ_0^2 の値 60.882 は棄却限界値 33.93031 より大きいので帰無仮説 $\mathcal{H}_0$: $\sigma^2 = 0.2$ は有意水準 0.05 で棄却されない。

4.3　対になった標本に関する検定

個体に対する薬の投与後の変化、学習を行った後での試験結果などの適用前の結果と適用後の結果を比べたいことがある。このデータは各個体に対応したデータで、適用前と適用後の差として表される。



（平均が増加した例）

適用後の差に関心があるので、各個体の観測値の差を $X_1, X_2, \cdots, X_n$ とし、これは平均 μ、分散 σ^2 の正規分布に従っていると仮定する。このとき、分散が既知の場合と未知の場合について以下の検定問題が考えられる。

両側検定　対立仮説 $\mathcal{H}_1\colon \mu \neq \mu_0$ に対して帰無仮説 $\mathcal{H}_0\colon \mu = \mu_0$ を検定する問題

片側検定 $\begin{cases} \text{対立仮説 } \mathcal{H}_1\colon \mu > \mu_0 \text{ に対して帰無仮説 } \mathcal{H}_0\colon \mu = \mu_0 \text{ を検定する問題} \\ \text{対立仮説 } \mathcal{H}_1\colon \mu < \mu_0 \text{ に対して帰無仮説 } \mathcal{H}_0\colon \mu = \mu_0 \text{ を検定する問題} \end{cases}$

$\mu_0 = 0$ としたとき、両側検定はなんらかの差があるということを調べたいとき、片側検定は各々、正の差がある、負の差があるということを調べたいときに用いることになる。

分散が既知のときは節 4.1.1(32 ページ) の結果を用い、分散が未知のときは節 4.1.2(35 ページ) の結果を用いれば良いことが分かる。つまり、

4.3.1　分散が既知のとき

統計量

$$Z = \frac{\overline{X} - \mu_0}{\sigma / \sqrt{n}}$$

の値を求める。ここで、

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

である。有意水準を α とし、z_α を正規分布の上側 α% 点[*5]とすると、棄却領域は下記のようになる。

帰無仮説 $\mathcal{H}_0\colon \mu = \mu_0$ *versus* 対立仮説 $\mathcal{H}_1\colon \mu \neq \mu_0$ $\begin{cases} |Z| > z_{\alpha/2} \text{ ならば帰無仮説を棄却する。} \\ |Z| \leq z_{\alpha/2} \text{ ならば帰無仮説は棄却できない。} \end{cases}$

帰無仮説 $\mathcal{H}_0\colon \mu = \mu_0$ *versus* 対立仮説 $\mathcal{H}_1\colon \mu > \mu_0$ $\begin{cases} Z > z_\alpha \text{ ならば帰無仮説を棄却する。} \\ Z \leq z_\alpha \text{ ならば帰無仮説は棄却できない。} \end{cases}$

帰無仮説 $\mathcal{H}_0\colon \mu = \mu_0$ *versus* 対立仮説 $\mathcal{H}_1\colon \mu < \mu_0$ $\begin{cases} Z < -z_\alpha \text{ ならば帰無仮説を棄却する。} \\ Z \geq -z_\alpha \text{ ならば帰無仮説は棄却できない。} \end{cases}$

4.3.2　分散が未知のとき

統計量

$$t = \frac{\overline{X} - \mu_0}{s / \sqrt{n}}$$

の値を求める。ここで

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2$$

である。有意水準を α とし、$t_{n-1}(\alpha)$ を自由度 $n-1$ の t 分布の上側 α% 点[*6]とすると、棄却領域は下記のようになる。

[*5] 33 ページ参照

[*6] 35 ページ参照。

帰無仮説 $\mathcal{H}_0$: $\mu = \mu_0$ *versus* 対立仮説 $\mathcal{H}_1$: $\mu \neq \mu_0$ $\begin{cases} |t| > t_{n-1}(\alpha/2) \text{ ならば帰無仮説を棄却する。} \\ |t| \leq t_{n-1}(\alpha/2) \text{ ならば帰無仮説は棄却できない。} \end{cases}$

帰無仮説 $\mathcal{H}_0$: $\mu = \mu_0$ *versus* 対立仮説 $\mathcal{H}_1$: $\mu > \mu_0$ $\begin{cases} t > t_{n-1}(\alpha) \text{ ならば帰無仮説を棄却する。} \\ t \leq t_{n-1}(\alpha) \text{ ならば帰無仮説は棄却できない。} \end{cases}$

帰無仮説 $\mathcal{H}_0$: $\mu = \mu_0$ *versus* 対立仮説 $\mathcal{H}_1$: $\mu < \mu_0$ $\begin{cases} t < -t_{n-1}(\alpha) \text{ ならば帰無仮説を棄却する。} \\ t \geq -t_{n-1}(\alpha) \text{ ならば帰無仮説は棄却できない。} \end{cases}$

例題 4.5. 10 人について入浴前の血圧と入浴 30 分後の血圧を測定しところ下記の表を得た。

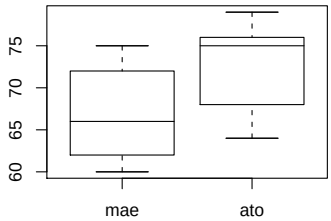
入浴前	61	69	68	72	62	60	63	75	64	72
入浴後	68	66	75	75	77	71	75	76	64	79

有意水準 5% で血圧の上昇が見られるかを検定せよ。■

入浴 30 分後の血圧と入浴前の血圧の差が平均 μ、分散未知の正規分布に従っていると仮定し、

帰無仮説 $\mathcal{H}_0$: $\mu = 0$ *versus* 対立仮説 $\mathcal{H}_1$: $\mu > 0$

を有意水準 5% で検定すればよい。二つの標本の比較のために箱ひげ図を書くと下記のようになる。



入力と出力結果

```
> b<- c(61, 69, 68, 72, 62, 60, 63, 75, 64, 72)   # 入浴前の血圧。
> a<- c(68, 66, 75, 75, 77, 71, 75, 76, 64, 79)   # 入浴後の血圧。
> x<- a-b                                         # 血圧差。
> t.test(x, mu=0, alternative="greater")          # t-検定。
        One Sample t-test                         # 一標本検定。
data:  x                                          # データ名：x。
t = 3.3085, df = 9, p-value = 0.004553            # t 値、自由度、p-値。
alternative hypothesis: true mean                 # 対立仮説：真の平均はゼロより大きい。
             is greater than 0
95 percent confidence interval:                   # 95% 信頼区間。
 2.675596      Inf                                # (2.675596, +∞)
sample estimates:                                 # 標本推定値。
mean of x                                         # x の平均。
        6
```

問題

問題 4.1.

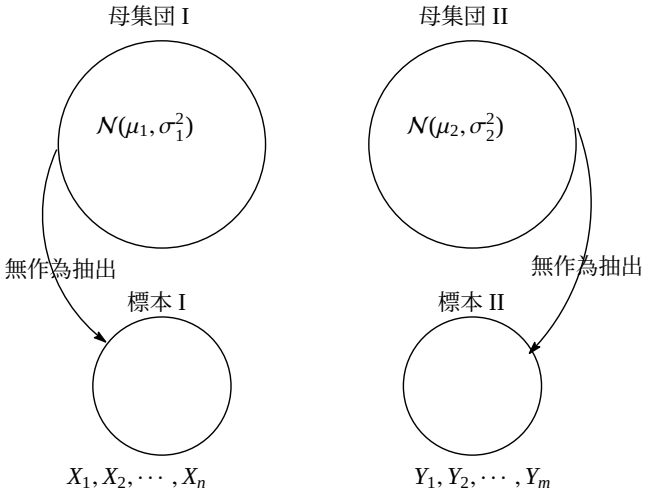
データ `iris` の種 `versicolor` のガク片の長さについて、有意水準 0.05 で以下の検定問題を行え。

1. 帰無仮説 $\mathcal{H}_0$: $\mu = 5.006$ *versus* 対立仮説 $\mathcal{H}_1$: $\mu > 5.006$ (分散：未知)
2. 帰無仮説 $\mathcal{H}_0$: $\sigma^2 = 0.1242490$ *versus* 対立仮説 $\mathcal{H}_1$: $\sigma^2 > 0.1242490$ (平均：未知)

ここで、5.006 と 0.1242490 は、それぞれ種 `setosa` の想定された平均と分散である。

第 5 章　二標本正規母集団に関する検定

二つの母集団が正規分布 (normal distribution) に従うと仮定する。この二つの母集団の平均と分散をそれぞれ、μ_i と σ_i^2 で表す ($i = 1, 2$)。母集団 I から大きさ n の標本 $X_1, X_2, \cdots, X_n$ が抽出され、これとは独立に母集団 II から大きさ m の標本 $Y_1, Y_2, \cdots, Y_m$ が抽出されている。



次の平均の差と分散の比についての検定問題を考える。

平均の差の検定

- 分散 σ_1^2 と σ_2^2 が共に既知
  - (1) $\mathcal{H}_0 : \mu_1 - \mu_2 = \delta_0$ *versus* $\mathcal{H}_1 : \mu_1 - \mu_2 \neq \delta_0$
  - (2) $\mathcal{H}_0 : \mu_1 - \mu_2 = \delta_0$ *versus* $\mathcal{H}_1 : \mu_1 - \mu_2 < \delta_0$
  - (3) $\mathcal{H}_0 : \mu_1 - \mu_2 = \delta_0$ *versus* $\mathcal{H}_1 : \mu_1 - \mu_2 > \delta_0$
- 分散 σ_1^2 と σ_2^2 は未知であるが等しいことが分かっている
  - (1) $\mathcal{H}_0 : \mu_1 - \mu_2 = \delta_0$ *versus* $\mathcal{H}_1 : \mu_1 - \mu_2 \neq \delta_0$
  - (2) $\mathcal{H}_0 : \mu_1 - \mu_2 = \delta_0$ *versus* $\mathcal{H}_1 : \mu_1 - \mu_2 < \delta_0$
  - (3) $\mathcal{H}_0 : \mu_1 - \mu_2 = \delta_0$ *versus* $\mathcal{H}_1 : \mu_1 - \mu_2 > \delta_0$
- 分散 σ_1^2 と σ_2^2 が共に未知
  - (1) $\mathcal{H}_0 : \mu_1 - \mu_2 = \delta_0$ *versus* $\mathcal{H}_1 : \mu_1 - \mu_2 \neq \delta_0$
  - (2) $\mathcal{H}_0 : \mu_1 - \mu_2 = \delta_0$ *versus* $\mathcal{H}_1 : \mu_1 - \mu_2 < \delta_0$
  - (3) $\mathcal{H}_0 : \mu_1 - \mu_2 = \delta_0$ *versus* $\mathcal{H}_1 : \mu_1 - \mu_2 > \delta_0$

分散の比の検定

- 平均 μ_1 と μ_2 が共に未知
  - (1) $\mathcal{H}_0 : \dfrac{\sigma_1^2}{\sigma_2^2} = 1$ *versus* $\mathcal{H}_1 : \dfrac{\sigma_1^2}{\sigma_2^2} \neq 1$
  - (2) $\mathcal{H}_0 : \dfrac{\sigma_1^2}{\sigma_2^2} = 1$ *versus* $\mathcal{H}_1 : \dfrac{\sigma_1^2}{\sigma_2^2} < 1$
  - (3) $\mathcal{H}_0 : \dfrac{\sigma_1^2}{\sigma_2^2} = 1$ *versus* $\mathcal{H}_1 : \dfrac{\sigma_1^2}{\sigma_2^2} > 1$

上記以外の検定問題が考えられるが、ここでは説明を省く。

5.1　平均の差の検定

5.1.1　分散 σ_1^2 と σ_2^2 が共に既知の場合

平均の差 $\mu_1 - \mu_2 = \delta_0$ は、母集団平均の差

$$\overline{X} - \overline{Y} = \frac{1}{n}\sum_{i=1}^{n} X_i - \frac{1}{m}\sum_{j=1}^{m} Y_j$$

によって推定され、この差の分散は、母集団 I と II の独立性から

$$\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}$$

である[*1]ので、この検定問題の検定統計量として、

$$\begin{aligned} Z &= \frac{(\overline{X} - \overline{Y}) - (\mu_1 - \mu_2)}{\sqrt{\dfrac{\sigma_1^2}{n} + \dfrac{\sigma_2^2}{m}}} \\ &= \frac{\overline{X} - \overline{Y} - \delta_0}{\sqrt{\dfrac{\sigma_1^2}{n} + \dfrac{\sigma_2^2}{m}}} \end{aligned} \tag{5.1}$$

を考えると、Z は帰無仮説の下で標準正規分布に従う。有意水準を α とするとき、下記のような棄却領域が作られる。

帰無仮説 $\mathcal{H}_0 : \mu_1 - \mu_2 = \delta_0$ *versus* 対立仮説 $\mathcal{H}_1 : \mu_1 - \mu_2 \neq \delta_0$ $\begin{cases} |Z| > z_{\alpha/2} \text{ ならば帰無仮説を棄却する。} \\ |Z| \leq z_{\alpha/2} \text{ ならば帰無仮説は棄却できない。} \end{cases}$

帰無仮説 $\mathcal{H}_0 : \mu_1 - \mu_2 = \delta_0$ *versus* 対立仮説 $\mathcal{H}_1 : \mu_1 - \mu_2 > \delta_0$ $\begin{cases} Z > z_{\alpha} \text{ ならば帰無仮説を棄却する。} \\ Z \leq z_{\alpha} \text{ ならば帰無仮説は棄却できない。} \end{cases}$

帰無仮説 $\mathcal{H}_0 : \mu_1 - \mu_2 = \delta_0$ *versus* 対立仮説 $\mathcal{H}_1 : \mu_1 - \mu_2 < \delta_0$ $\begin{cases} Z < -z_{\alpha} \text{ ならば帰無仮説を棄却する。} \\ Z \geq -z_{\alpha} \text{ ならば帰無仮説は棄却できない。} \end{cases}$

例題 5.1. R のデータ `ToothGrowth` を考える。これはモルモットの歯の成長 (`length`) と摂取方法 (`supp`) についてのデータである。`supp` は `OJ`(orange juice) と `VC`(Vitamin C) からなり、過去の経験から `OJ` と `VC` に対する母分散はそれぞれ、68.3 と 43.6 であることが分かっているものとする。`OJ` と `VC` の平均の差についての仮説検定を有意水準 5% で考える。

二群のヒストグラムと箱ひげ図は図 5.1 のようになっている。[*2]。■

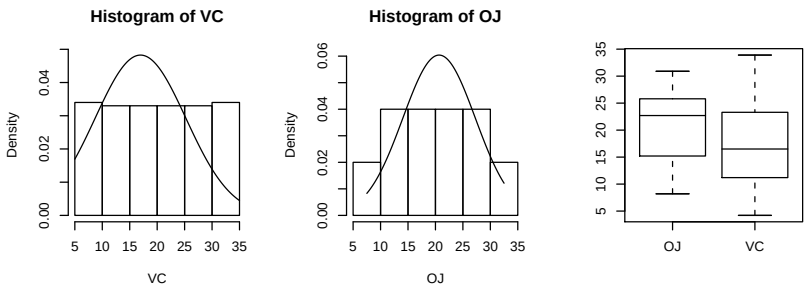


図 5.1　データ `ToothGrowth` のヒストグラムと箱ひげ図

[*1] 正規分布の再生性による。

[*2] 二群の正規性はいまは問わない。コマンド `example(ToothGrowth)` を実行すると別の解析法が示されていることが分かる。

(I) 両側検定: 帰無仮説 $\mathcal{H}_0$: $\mu_1 - \mu_2 = -3.5$ *versus* 対立仮説 $\mathcal{H}_1$: $\mu_1 - \mu_2 \neq -3.5$ (分散は既知で 68.3 と 43.6 と仮定)
検定統計量

$$Z = \frac{\overline{X} - \overline{Y} - \delta_0}{\sqrt{\dfrac{\sigma_1^2}{n} + \dfrac{\sigma_2^2}{m}}}$$

の絶対値が $z_{0.025} = 1.96$ より大きいとき帰無仮説を棄却する。

入力と出力結果

```
> data(ToothGrowth)
> ToothGrowth                                  # データ ToothGrowth の内容表示。
> help(ToothGrowth)                            # データ ToothGrowth の説明表示。
> boxplot(len~supp, data=ToothGrowth)          # 箱ひげ図。
> x<- ToothGrowth[ToothGrowth$supp=="VC",1]    # x を摂取方法が VC のもの。
> y<- ToothGrowth[ToothGrowth$supp=="OJ",1]    # y を摂取方法が OJ のもの。
> n<- length(x)                                # x のデータ数。
> m<- length(y)                                # y のデータ数。
> (mean(x)-mean(y)+3.5)/sqrt(68.3/n+43.6/m)    # 検定統計量の計算。
[1] -0.1035561
```

検定統計量の値の絶対値は 0.1035561 で、$z_{0.025} = 1.96$ と比べて小さいので、有意水準 0.05 で帰無仮説が棄却されない。

(II) 片側検定: 帰無仮説 $\mathcal{H}_0$: $\mu_1 - \mu_2 = -3$ *versus* 対立仮説 $\mathcal{H}_1$: $\mu_1 - \mu_2 > -3$ (分散は既知で 68.3 と 43.6 と仮定)
検定統計量

$$Z = \frac{\overline{X} - \overline{Y} - \delta_0}{\sqrt{\dfrac{\sigma_1^2}{n} + \dfrac{\sigma_2^2}{m}}}$$

の値が $z_{0.05} = 1.64$ より大きいとき帰無仮説を棄却する。

入力と出力結果

```
> (mean(x)-mean(y)+3)/sqrt(68.3/n+43.6/m)    # 検定統計量の計算。
[1] -0.3624463
```

検定統計量の値は −0.3624463 で、$z_{0.05} = 1.64$ と比べて小さいので、有意水準 0.05 で帰無仮説が棄却されない。

(III) 片側検定: 帰無仮説 $\mathcal{H}_0$: $\mu_1 - \mu_2 = 0$ *versus* 対立仮説 $\mathcal{H}_1$: $\mu_1 - \mu_2 < 0$ (分散は既知で 68.3 と 43.6 と仮定)
検定統計量

$$Z = \frac{\overline{X} - \overline{Y} - \delta_0}{\sqrt{\dfrac{\sigma_1^2}{n} + \dfrac{\sigma_2^2}{m}}}$$

の値が $-z_{0.05} = -1.64$ より小さいとき帰無仮説を棄却する。

入力と出力結果

```
> (mean(x)-mean(y))/sqrt(68.3/n+43.6/m)    # 検定統計量の計算。
[1] -1.915787
```

検定統計量の値は −1.915787 で、$-z_{0.025} = -1.96$ と比べて大きいので、有意水準 0.05 で帰無仮説が棄却されない。

5.1.2 分散 σ_1^2 と σ_2^2 は未知であるが等しいことが分かっている場合

二つの母集団の母分散 σ_1^2 と σ_2^2 の値は未知であるが、同じ散らばりをもつ状況を考える。これは新しい機械からの製品は従来の機械と変わらない精度 (散らばり) を持つことが想定できたり、二群でそれほどの違いはなく平均のみ違っていると考えられるときで

ある。このとき、二つの分散が等しいと仮定しているので、標本 $X_1, X_2, \cdots, X_n$ と $Y_1, Y_2, \cdots, Y_m$ をまとめて (プールして)、分散を

$$\begin{aligned} V &= \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2 + \sum_{j=1}^{m}(Y_j - \overline{Y})^2}{n+m-2} \\ &= \frac{(n-1)s_1^2 + (m-1)s_2^2}{n+m-2} \end{aligned}$$

によって推定する。ここで、

$$s_1^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2, \qquad s_2^2 = \frac{1}{m-1}\sum_{j=1}^{m}(Y_j - \overline{Y})^2$$

である。式 (5.1) と同じような考え方により、ここでの検定問題に対して検定統計量

$$t = \frac{\overline{X} - \overline{Y} - \delta_0}{\sqrt{(\frac{1}{n} + \frac{1}{m})V}}$$

が導かれる。このとき、t は自由度 $n+m-2$ の t 分布に従うことが知られており、有意水準 α に対して下記のような棄却領域が作られる。[*3]。

帰無仮説 $\mathcal{H}_0 : \mu_1 - \mu_2 = \delta_0$ *versus* 対立仮説 $\mathcal{H}_1$: $\mu_1 - \mu_2 \neq \delta_0$ $\begin{cases} |t| > t_{n+m-2}(\alpha/2) \text{ ならば帰無仮説を棄却する。} \\ |t| \leq t_{n+m-2}(\alpha/2) \text{ ならば帰無仮説は棄却できない。} \end{cases}$

帰無仮説 $\mathcal{H}_0 : \mu_1 - \mu_2 = \delta_0$ *versus* 対立仮説 $\mathcal{H}_1$: $\mu_1 - \mu_2 > \delta_0$ $\begin{cases} t > t_{n+m-2}(\alpha) \text{ ならば帰無仮説を棄却する。} \\ t \leq t_{n+m-2}(\alpha) \text{ ならば帰無仮説は棄却できない。} \end{cases}$

帰無仮説 $\mathcal{H}_0 : \mu_1 - \mu_2 = \delta_0$ *versus* 対立仮説 $\mathcal{H}_1$: $\mu_1 - \mu_2 < \delta_0$ $\begin{cases} t < -t_{n+m-2}(\alpha) \text{ ならば帰無仮説を棄却する。} \\ t \geq -t_{n+m-2}(\alpha) \text{ ならば帰無仮説は棄却できない。} \end{cases}$

例題 5.2. R のデータ `ToothGrowth` を考え、`OJ` と `VC` の平均の差についての仮説検定を扱う。ただし、各分散の値は未知であるが、等しいと仮定できるものとし、有意水準は 5% とする。■

(I) 両側検定: 帰無仮説 $\mathcal{H}_0$: $\mu_1 - \mu_2 = -3.5$ *versus* 対立仮説 $\mathcal{H}_1$: $\mu_1 - \mu_2 \neq -3.5$ (分散は未知であるが、等しいと仮定)
検定統計量

$$t = \frac{\overline{X} - \overline{Y} - \delta_0}{\sqrt{(1/n + 1/m)V}}$$

の絶対値が $t_{n+m-2}(\alpha/2)$ より大きいとき帰無仮説を棄却する。

入力と出力結果

```
> x<- ToothGrowth[ToothGrowth$supp=="VC",1]                    # x を摂取方法が VC のもの。
> y<- ToothGrowth[ToothGrowth$supp=="OJ",1]                    # y を摂取方法が OJ のもの。
> n<- length(x); m<- length(y)                                 # データ数を n、m とする。
 > t.test(x, y, mu=-3.5, var.equal=T, alternative="two.sided")  # 分散は等しいと仮定。
        Two Sample t-test                                      # 二標本 t-検定。
data:  x and y                                                 # データ名: x と y。
t = -0.1035, df = 58, p-value = 0.9179                         # t 値、自由度、p 値。
alternative hypothesis: true difference in means               # 対立仮説：平均の差は-3.5 でない。
           is not equal to -3.5
95 percent confidence interval:                                # 95% 信頼区間。
 -7.5670064  0.1670064
sample estimates:                                              # 標本推定値。
mean of x mean of y                                            # x の平均、y の平均。
 16.96333  20.556333
>  V<- ((n-1)*var(x)+(m-1)*var(y))/(n+m-2)                     # プールした標本の分散。
>  (mean(x)-mean(y)+3.5)/sqrt((1/n+1/m)*V)                     # 検算。
[1] -0.1035280                                                 # 上の結果と一致
```

[*3] t 分布については 35 ページ参照。

この検定問題での自由度は 58 で棄却限界値は `qt(.975,df=58)=2.001717` で、t-値は -0.1035 であるので帰無仮説は有意水準 5% で棄却されない。また、上で示される結果より、p 値は `0.9179` であるから有意水準 0.05 で帰無仮説が棄却されないことが分かる。

(II) 片側検定: 帰無仮説 $\mathcal{H}_0$: $\mu_1 - \mu_2 = -3$ *versus* 対立仮説 $\mathcal{H}_1$: $\mu_1 - \mu_2 > -3$ (分散は未知であるが、等しいと仮定)
検定統計量

$$t = \frac{\overline{X} - \overline{Y} - \delta_0}{\sqrt{(1/n + 1/m)V}}$$

の値が $t_{n+m-2}(\alpha)$ より大きいとき帰無仮説を棄却する。

入力と出力結果

```
> t.test(x, y, mu=-3, var.equal=T, alternative="greater")   # 分散は等しいと仮定。
        Two Sample t-test                                   # 二標本 t-検定。
data:  x and y                                              # データ名：x と y。
t = -0.3623, df = 58, p-value = 0.64088                     # t 値、自由度、p 値。
alternative hypothesis: true difference in means
           is greater than -3
95 percent confidence interval:                             # 95% 信頼区間。
 -6.92918      Inf                                          (−6.92918, +∞)
sample estimates:                                           # 標本推定値。
mean of x mean of y                                         # x の平均、y の平均。
 16.96333  20.556333
> V<- ((n-1)*var(x)+(m-1)*var(y))/(n+m-2)                   # プールした標本の分散。
> (mean(x)-mean(y)+3)/sqrt((1/n+1/m)*V)                     # 検算。
[1] -0.3623481                                              # 上の結果と一致。
```

この検定問題での自由度は 58 で棄却限界値は `qt(.95,df=58)=1.671553` で、t-値は -0.3623481 であるので帰無仮説は有意水準 5% で棄却されない。また、上で示される結果より、p 値は `0.6408` であるから有意水準 0.05 で帰無仮説が棄却されないことが分かる。

(III) 片側検定: 帰無仮説 $\mathcal{H}_0$: $\mu_1 - \mu_2 = 0$ *versus* 対立仮説 $\mathcal{H}_1$: $\mu_1 - \mu_2 < 0$ (分散は未知であるが、等しいと仮定)
検定統計量

$$t = \frac{\overline{X} - \overline{Y} - \delta_0}{\sqrt{(1/n + 1/m)V}}$$

の値が $-t_{n+m-2}(\alpha)$ より小さいとき帰無仮説を棄却する。

入力と出力結果

```
> t.test(x, y, mu=0, var.equal=T, alternative="less")   # 分散は等しいと仮定。
        Two Sample t-test                                # 二標本 t-検定。
data:  x and y                                           # データ名：x と y。
t = -1.9153, df = 58, p-value = 0.03020
alternative hypothesis: true difference in means
           is less than 0
95 percent confidence interval:                          # 95% 信頼区間。
       -Inf -0.4708204                                   (−∞, −0.4708204)
sample estimates:                                        # 推定値。
mean of x mean of y                                      # x の平均、y の平均。
 16.96333  20.556333
> V<- ((n-1)*var(x)+(m-1)*var(y))/(n+m-2)                # プールした標本の分散。
> (mean(x)-mean(y))/sqrt((1/n+1/m)*V)                    # 検算。
[1] -1.915268                                            # 上の結果と一致。
```

この検定問題での自由度は 58 で棄却限界値は `qt(.05,df=58)=-1.671553` で、t-値は -1.915268 であるので帰無仮説は有意水準 5% で棄却される。また、上で示される結果より、p 値は `0.03020` であるから有意水準 0.05 で帰無仮説が棄却されることが分かる。

5.1.3　分散 σ_1^2 と σ_2^2 が共に未知の場合

分散 σ_1^2 と σ_2^2 が共に未知のときの検定手法は Welch(ウェルチ) の検定と呼ばれ、前節で用いた検定統計量

$$t = \frac{\overline{X} - \overline{Y} - \delta_0}{\sqrt{(\frac{1}{n} + \frac{1}{m})V}}$$

を評価する際の自由度を $n + m - 2$ でなく、

$$\phi = \frac{(s_1/n + s_2/m)^2}{\frac{(s_1/n)^2}{n-1} + \frac{(s_2/m)^2}{m-1}}$$

を用いるものである。有意水準 α に対して下記のような棄却領域が作られる。

帰無仮説 $\mathcal{H}_0 : \mu_1 - \mu_2 = \delta_0$ *versus* 対立仮説 $\mathcal{H}_1$: $\mu_1 - \mu_2 \neq \delta_0$ $\begin{cases} |t| > t_\phi(\alpha/2) \text{ ならば帰無仮説を棄却する。} \\ |t| \leq t_\phi(\alpha/2) \text{ ならば帰無仮説は棄却できない。} \end{cases}$

帰無仮説 $\mathcal{H}_0 : \mu_1 - \mu_2 = \delta_0$ *versus* 対立仮説 $\mathcal{H}_1$: $\mu_1 - \mu_2 > \delta_0$ $\begin{cases} t > t_\phi(\alpha) \text{ ならば帰無仮説を棄却する。} \\ t \leq t_\phi(\alpha) \text{ ならば帰無仮説は棄却できない。} \end{cases}$

帰無仮説 $\mathcal{H}_0 : \mu_1 - \mu_2 = \delta_0$ *versus* 対立仮説 $\mathcal{H}_1$: $\mu_1 - \mu_2 < \delta_0$ $\begin{cases} t < -t_\phi(\alpha) \text{ ならば帰無仮説を棄却する。} \\ t \geq -t_\phi(\alpha) \text{ ならば帰無仮説は棄却できない。} \end{cases}$

例題 5.3. R のデータ `ToothGrowth` を考え、平均の差についての仮説検定を扱う。ただし、各分散は未知で、有意水準は 5% とする。■

(I) 両側検定: 帰無仮説 $\mathcal{H}_0$: $\mu_1 - \mu_2 = -3.5$ *versus* 対立仮説 $\mathcal{H}_1$: $\mu_1 - \mu_2 \neq -3.5$ (分散は未知)
検定統計量

$$t = \frac{\overline{X} - \overline{Y} - \delta_0}{\sqrt{(1/n + 1/m)V}}$$

の絶対値が $t_\phi(\alpha/2)$ より大きいとき帰無仮説を棄却する。

入力と出力結果

```
> x<- ToothGrowth[ToothGrowth$supp=="VC",1]        # x を摂取方法が VC のもの。
> y<- ToothGrowth[ToothGrowth$supp=="OJ",1]        # y を摂取方法が OJ のもの。
> t.test(x, y, mu=-3.5, alternative="two.sided")   # 対立仮説は両側。
        Welch Two Sample t-test                    # ウェルチの検定。
data:  x and y                                     # データ名：x と y。
t = -0.1035, df = 55.309, p-value = 0.918          # t 値、自由度、p 値。
alternative hypothesis: true difference in means   # 対立仮説。
           is not equal to -3.5
95 percent confidence interval:                    # 95% 信頼区間。
 -7.5710156  0.1710156                             (−7.5710156, 0.1710156)
sample estimates:                                  # 標本推定値。
mean of x mean of y                                # x の平均、y の平均。
 16.96333  20.556333
```

p 値は `0.918` であるから有意水準 0.05 で帰無仮説が棄却されない。

(II) 片側検定: 帰無仮説 $\mathcal{H}_0$: $\mu_1 - \mu_2 = -3$ *versus* 対立仮説 $\mathcal{H}_1$: $\mu_1 - \mu_2 > -3$ (分散は未知)
検定統計量

$$t = \frac{\overline{X} - \overline{Y} - \delta_0}{\sqrt{(1/n + 1/m)V}}$$

の値が $t_\phi(\alpha)$ より大きいとき帰無仮説を棄却する。

入力と出力結果

```
> t.test(x, y, mu=-3, alternative="greater")      # 対立仮説は「大きい」。
        Welch Two Sample t-test                   # ウェルチの検定。
data:  x and y                                    # データ名：x と y。
t = -0.3623, df = 55.309, p-value = 0.55408       # t 値、自由度、p 値。
alternative hypothesis: true difference in means  # 対立仮説。
           is not equal to -3
95 percent confidence interval:                   # 95% 信頼区間。
 -6.931731       Inf                              (−6.931731, +∞)
sample estimates:                                 # 標本推定値。
mean of x mean of y                               # x の平均、y の平均。
 16.96333  20.556333
```

p 値は `0.55408` であるから有意水準 0.05 で帰無仮説が棄却されない。

(III) 片側検定: 帰無仮説 $\mathcal{H}_0$: $\mu_1 - \mu_2 = 0$ *versus* 対立仮説 $\mathcal{H}_1$: $\mu_1 - \mu_2 < 0$ (分散は未知)
検定統計量

$$t = \frac{\overline{X} - \overline{Y} - \delta_0}{\sqrt{(1/n + 1/m)V}}$$

の値が $-t_\phi(\alpha)$ より小さいとき帰無仮説を棄却する。

入力と出力結果

```
> t.test(x, y, mu=0, alternative="less")          # 対立仮説は「小さい」。
        Welch Two Sample t-test                   # ウェルチの検定。
data:  x and y                                    # データ名：x と y。
t = -1.9153, df = 55.309, p-value = 0.03032       # t 値、自由度、p 値。
alternative hypothesis: true difference in means  # 対立仮説。
           is less than 0
95 percent confidence interval:                   # 95% 信頼区間。
        -Inf -0.4682687                           (−∞, −0.4682687)
sample estimates:                                 # 標本推定値。
mean of x mean of y                               # x の平均、y の平均。
 16.96333  20.556333
```

p 値は `0.03032` であるから有意水準 0.05 で帰無仮説が棄却される。

5.2　分散の比 σ_1^2/σ_2^2 の検定

二つの母集団の散らばりの程度の違いを見るために分散の比 σ_1^2/σ_2^2 について注目する。検定問題を扱うとき、母平均について仮定が必要となるが、ここでは母平均 μ_1 と μ_2 が共に未知の場合だけを扱う。

5.2.1　母平均 μ_1 と μ_2 が共に未知の場合

各母分散 σ_i^2 $(i = 1, 2)$ の推定量はそれぞれ

$$s_1^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2, \qquad s_2^2 = \frac{1}{m-1}\sum_{j=1}^{m}(Y_j - \overline{Y})^2$$

であり、分散の比 σ_1^2/σ_2^2 を推定するには対応するものを置き換えた

$$F_0 = \frac{s_1^2}{s_2^2} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2/(n-1)}{\sum_{j=1}^{m}(Y_j - \overline{Y})^2/(m-1)}$$

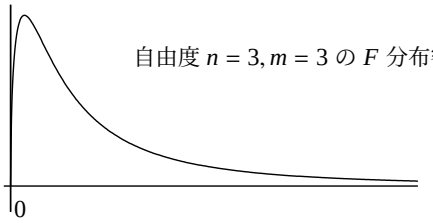
を用いればよく、この統計量 F_0 は自由度 $(n-1, m-1)$ の F 分布に従うことが知られている。ここで、自由度 (n, m) の F 分布関数は

$$\int_0^x \frac{\Gamma((n+m)/2)}{\Gamma(n/2)\Gamma(m/2)}\left(\frac{n}{m}\right)^{n/2}\frac{t^{(n-2)/2}}{[1 + (n/m)t]^{(n+m)/2}}\,dt$$

で与えられ、F 分布の上側 α% 点を

$$F(n,m,\alpha)=\int_\alpha^\infty \frac{\Gamma((n+m)/2)}{\Gamma(n/2)\Gamma(m/2)}\left(\frac{n}{m}\right)^{n/2}\frac{t^{(n-2)/2}}{[1+(n/m)t]^{(n+m)/2}}\,dt$$

によって定義する。ここで、F 分布密度関数は



自由度 (3,3) の F 分布密度関数

```
fdensity<- function(x) df(x,3,3)
curve(fdensity,0,3)
```

のような形をしている。

このとき、有意水準 α に対して、以下の棄却域が作られ、F **検定**(F test) と呼ばれる。

帰無仮説 $\mathcal{H}_0$: $\dfrac{\sigma_1}{\sigma_2}=1$ *versus* 対立仮説 $\mathcal{H}_1$: $\dfrac{\sigma_1}{\sigma_2}\neq 1$
$$\begin{cases} s_1^2>s_2^2,\ F_0>F(n-1,m-1,1-\alpha/2) \text{ ならば帰無仮説を棄却する。} \\ s_1^2<s_2^2,\ \dfrac{1}{F_0}>F(m-1,n-1,1-\alpha/2) \text{ ならば帰無仮説を棄却する。} \\ \text{上記以外ならば帰無仮説を棄却できない。} \end{cases}$$

帰無仮説 $\mathcal{H}_0$: $\dfrac{\sigma_1}{\sigma_2}=1$ *versus* 対立仮説 $\mathcal{H}_1$: $\dfrac{\sigma_1}{\sigma_2}>1$
$$\begin{cases} F_0>F(n-1,m-1,1-\alpha) \text{ ならば帰無仮説を棄却する。} \\ F_0\leq F(n-1,m-1,1-\alpha) \text{ ならば帰無仮説は棄却できない。} \end{cases}$$

帰無仮説 $\mathcal{H}_0$: $\dfrac{\sigma_1}{\sigma_2}=1$ *versus* 対立仮説 $\mathcal{H}_1$: $\dfrac{\sigma_1}{\sigma_2}<1$
$$\begin{cases} \dfrac{1}{F_0}<F(m-1,n-1,1-\alpha) \text{ ならば帰無仮説を棄却する。} \\ \dfrac{1}{F_0}\geq F(m-1,n-1,1-\alpha) \text{ ならば帰無仮説は棄却できない。} \end{cases}$$

例題 5.4. R のデータ `ToothGrowth` を考え、分散の比についての仮説検定を扱う。各平均は未知とし、有意水準は 5% とする。■

(I) 帰無仮説 $\mathcal{H}_0$: $\dfrac{\sigma_1}{\sigma_2}=1$ *versus* 対立仮説 $\mathcal{H}_1$: $\dfrac{\sigma_1}{\sigma_2}\neq 1$ (各平均は未知)

各分散の推定量 s_1^2、s_2^2 に対して、

$$s_1^2>s_2^2, F_0=\frac{s_1^2}{s_2^2}>F(n-1,m-1,0.975)\Rightarrow \text{帰無仮説を棄却する。}$$
$$s_1^2<s_2^2, \frac{1}{F_0}=\frac{s_2^2}{s_1^2}>F(m-1,n-1,0.975)\Rightarrow \text{帰無仮説を棄却する。}$$
$$\text{上記以外のとき}\Rightarrow \text{帰無仮説は棄却できない。}$$

という検定方式である。

入力と出力結果

```
> data(ToothGrowth)                                    # データ ToothGrowth の組み込み。
> x<- ToothGrowth[ToothGrowth$supp=="VC",1]            # x を摂取方法が VC のもの。
> y<- ToothGrowth[ToothGrowth$supp=="OJ",1]            # y を摂取方法が OJ のもの。
> var.test(x, y, alternative="two.sided")              # 対立仮説は両側。
        F test to compare two variances                # 二つの分散を比較するための F 検定。
data:  x and y                                         # データ名：x と y。
F = 1.5659, num df = 29, denom df = 29, p-value = 0.2331  # F0 の値、分子、分母の自由度 p 値。
alternative hypothesis: true ratio of variances        # 対立仮説：真の分散比は 1 でない。
           is not equal to 1
95 percent confidence interval:                        # 95% 信頼区間。
 0.745331 3.290028
sample estimates:                                      # 標本推定値。
ratio of variances                                     # 分散比。
          1.565937
```

p 値は `0.2331` であるので、有意水準 0.05 では帰無仮説が棄却されない。

(II) 帰無仮説 $\mathcal{H}_0$: $\dfrac{\sigma_1}{\sigma_2} = 1$ *versus* 対立仮説 $\mathcal{H}_1$: $\dfrac{\sigma_1}{\sigma_2} > 1$ (各平均は未知)

各分散の推定量 s_1^2、s_2^2 に対して、

$$F_0 = \frac{s_1^2}{s_2^2} > F(n-1, m-1, 0.95) \Rightarrow \text{帰無仮説を棄却する。}$$

$$\text{上記以外のとき} \Rightarrow \text{帰無仮説は棄却できない。}$$

という検定方式であった。

入力と出力結果

```
> var.test(x, y, alternative="greater")                    # 対立仮説は「大きい」。
        F test to compare two variances                    # 二つの分散を比較するための F 検定。
data:  x and y                                             # データ名：x と y。
F = 1.5659, num df = 29, denom df = 29, p-value = 0.1166   # F0 の値、分子、分母の自由度 p 値。
alternative hypothesis: true ratio of variances            # 対立仮説：真の分散比は 1 より大きい。
           is greater than 1
95 percent confidence interval:                            # 95% 信頼区間。
 0.8415347       Inf                                       [0.8415347, +∞)
sample estimates:                                          # 標本推定値。
ratio of variances                                         # 分散比。
          1.565937
```

p 値は `0.1166` であるので、有意水準 0.05 では帰無仮説が棄却されない。

(III) 帰無仮説 $\mathcal{H}_0$: $\dfrac{\sigma_1}{\sigma_2} = 1$ *versus* 対立仮説 $\mathcal{H}_1$: $\dfrac{\sigma_1}{\sigma_2} < 1$ (各平均は未知)

各分散の推定量 s_1^2、s_2^2 に対して、

$$\frac{1}{F_0} = \frac{s_2^2}{s_1^2} > F(m-1, n-1, 0.95) \Rightarrow \text{帰無仮説を棄却する。}$$

$$\text{上記以外のとき} \Rightarrow \text{帰無仮説は棄却できない。}$$

という検定方式であった。よって、

```
var.test(x, y, alternative="less")
```

というコマンドでなく、

```
var.test(y, x, alternative="greater")
```

を用いることになります。

入力と出力結果

```
> var.test(y, x, alternative="greater")                    # 対立仮説は「大きい」。
        F test to compare two variances                    # 二つの分散を比較するための F 検定。
data:  y and x                                             # データ名：x と y
0.6386, num df = 29, denom df = 29, p-value =  0.8834      # F0 の値、分子、分母の自由度 p 値。
alternative hypothesis: true ratio of variances            # 対立仮説：真の分散比は 1 より大きい。
           is greater than 1
95 percent confidence interval:                            # 95% 信頼区間。
 0.343181      Inf                                         [0.343181, ∞)
sample estimates:                                          # 標本推定値。
ratio of variances                                         # 分散比。
          0.6385951
```

p 値は `0.8834` であるので、有意水準 0.05 では帰無仮説が棄却さない。検定統計量の値は、0.6386 で、自由度 (29, 29) の F 分布の上側 5% 点は 1.860811 である。ここで、自由度 (29, 29) の F 分布の上側 5% 点 $F(0.95, 29, 29)$ は、`qf(0.95,29,29)=1.860811` で与えられる。

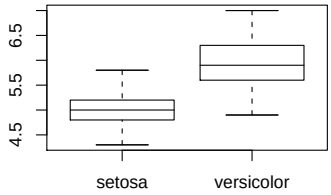
問題

問題 5.1.

データ `iris` の種 `setosa` と `versicolor` のガク片の長さについて、有意水準 0.05 で

$$\text{帰無仮説 } \mathcal{H}_0\colon \mu_1 = \mu_2 \textit{ versus } \text{対立仮説 } \mathcal{H}_1\colon \mu_1 < \mu_2$$

を検定せよ。ただし、各分散は未知であるとする。ここで、μ_1、μ_2 はそれぞれ、`setosa` と `versicolr` のガク片の平均を表す。

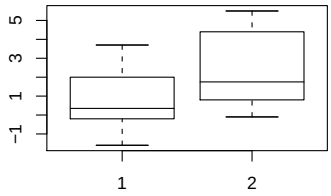


問題 5.2.

データ `sleep` の `group` ごとの `extra` の平均に差があるかないかを判断するために以下のことを考察せよ。

(i) 分散の比の検定を行う。
(ii) 分散が等しいと仮定し、平均の差の検定を行う。

ただし、両検定とも有意水準は 0.05 とする。



Guinea Pigs

第 6 章　適合度の検定

6.1　度数の適合性

試行結果が h 個の階級 $A_i, (i = 1, 2, \cdots, h)$ に分類される実験を n 回繰り返したとき、各階級 A_i の発生度数を f_i で表すと

$$f_1 + f_2 + \cdots + f_h = n$$

である。

階級	A_1	A_2	$\cdots$	A_h	計
観測頻度	f_1	f_2	$\cdots$	f_h	n
確率	p_1	p_2	$\cdots$	p_h	1
期待頻度	np_1	np_2	$\cdots$	np_h	n

このとき、

帰無仮説 $\mathcal{H}_0$: 各階級の発生確率は p_i である

を検定することを考える。ここで

$$\sum_{i=1}^{h} p_i = 1, \qquad 0 \le p_i \le 1$$

である。帰無仮説の下で期待される階級の度数は

$$(np_1, np_2, \cdots, np_h)$$

であり、実際観測された度数

$$(f_1, f_2, \cdots, f_h)$$

との差を測るものとして、

$$\chi^2 = \sum_{i=1}^{h} \frac{(f_i - np_i)^2}{np_i}$$

を考える。これは各階級 i の観測度数から理論度数を引いたものを二乗し、np_i で割ったものの総和となっている。この考え方は K. Pearson によるもので、統計量 χ^2 は近似的に自由度 $h-1$ のカイ二乗分布[*1]に従うことが知られている。

例題 6.1. 平成 12 年度の 0 歳児の出生月 (4 区分) ごとの総数は

1 月〜3 月	4 月〜6 月	7 月〜9 月	10 月〜12 月
151,422	149,233	152,455	147,356

であった[*2]。4 区分において出生数に差があるか否かを有意水準 5% で検定せよ。■

入力と出力結果

```
> o<- c(151422, 149233, 152455, 147356)          # 観測度数を o とする。
> prob<- c(1,1,1,1)/4                            # 理論確率を prob とする。
> chisq.test(o,p=prob)                           # o と prob を用いたカイ二乗検定。
        Chi-squared test for given probabilities  # カイ二乗検定。
data:  o                                         # データ名は o。
X-squared = 103.7451, df = 3, p-value < 2.2e-16  # カイ二乗値、自由度、p 値。
```

[*1] 38 ページ参照。

[*2] http://www.stat.go.jp/data/kokusei/2000/kihon1/00/hyodai.htm からのデータである。

p 値が 2.2e-16 以下であるので、4 区分での出生数が同じであるという仮説の下で、上のような観測値が発生する確率はゼロに近いことが分かる (自由度 3 のカイ二乗分布の上側 5% 点は qchisq(0.95,df=3)=7.814728 である)。よって、帰無仮説：「4 区分で出生数が同じである」は有意水準 5% で棄却される。また、検算として、カイ二乗値は右のコマンドによっても得られる。

入力と出力結果

```
> o<- c(151422, 149233, 152455, 147356)
> e<- c(1,1,1,1)*sum(o)/4
> sum((o-e)^2/e)
[1] 103.7451
```

例題 6.2. メンデルの遺伝の法則を考える。実験データはエンドウ豆についての色 (黄色・緑色) とかたち (丸い・しわ) で分類した観測度数が示されている。メンデルの法則が有意水準 5% で成り立つか検定せよ。■

豆の種類	黄色・丸い	黄色・しわ	緑色・丸い	緑色・しわ	計
観測度数	315	101	108	32	556
メンデルの法則	$\frac{9}{16}$	$\frac{3}{16}$	$\frac{3}{16}$	$\frac{1}{16}$	1

入力と出力結果

```
> o<- c(315, 108, 101, 32 )                    # 観測度数を o とする。
> prob<- c(9,3,3,1)/16                         # 理論確率を p とする。
> chisq.test(o,p=prob)                         # o と prob を用いたカイ二乗検定。
        Chi-squared test for given probabilities  # カイ二乗検定。
data:  o                                       # データ名は o。
X-squared = 0.47, df = 3, p-value = 0.9254     # カイ二乗値、自由度、p 値。
```

p 値が 0.9254 であるので、有意水準 5% で帰無仮説は棄却されない。この場合のカイ二乗値は 0.47 で、自由度 3 のカイ二乗分布の下側 5% 点は qchisq(0.05,df=3)=0.3518463 であるので、帰無仮説は棄却されない。また、検算として、カイ二乗値は右のコマンドによっても得られる。

入力と出力結果

```
> o<- c(315, 108, 101, 32 )
> e<- c(9,3,3,1)*sum(o)/16
> sum((o-e)^2/e)
[1] 0.470024
```

6.2　分割表と独立性の検定

n 個の標本が 2 つの属性 A, B に関して分類されていると仮定する。ただし、属性 A は h 個の水準 $A_1, A_2, \cdots, A_h$ に細分され、属性 B は k 個の水準 $B_1, B_2, \cdots, B_k$ に細分されているものとする。また、各水準の組み合わせ (A_i, B_j) に属する観測度数は f_{ij} であるとする。このようにして出来る右図のような表を $h \times k$ **分割表**($h \times k$ contingency table) という。

	B_1	B_2	$\cdots$	B_k	計
A_1	f_{11}	f_{12}	$\cdots$	f_{1k}	$n_{1\cdot}$
A_2	f_{21}	f_{22}	$\cdots$	f_{2k}	$n_{2\cdot}$
$\vdots$	$\cdots$	$\cdots$	$\ddots$	$\cdots$	$\vdots$
A_h	f_{h1}	f_{h2}	$\cdots$	f_{hk}	$n_{h\cdot}$
計	$n_{\cdot 1}$	$n_{\cdot 2}$	$\cdots$	$n_{\cdot k}$	n

右図の表において、

$$n_{\cdot j} = \sum_{i=1}^{h} f_{ij}, \qquad n_{i\cdot} = \sum_{j=1}^{k} f_{ij},$$

$$n = \sum_{i=1}^{h} n_{i\cdot} = \sum_{j=1}^{k} n_{\cdot j} = \sum_{i=1}^{h}\sum_{j=1}^{k} f_{ij}$$

である。ここで、$n_{i\cdot}$ を第 i 行周辺度数と呼び、$n_{\cdot j}$ を第 j 列周辺度数と呼び、。

二つの属性 A と B に関連性がない、つまり統計的に独立という状況を考える。属性 A に依存しなく B の水準 B_j 起こる確率 $\Pr\{B_j\}$ が一定で、同じように属性 B に依存しなく A の水準 A_j 起こる確率 $\Pr\{A_j\}$ が一定であるので、A_i と B_j が同時に起こる確率 $\Pr\{A_iB_j\}$ に対して

$$\Pr\{A_iB_j\} = \Pr\{A_i\}\Pr\{B_j\}$$

という関係式が成立する。属性 A と B 間に関連がないという帰無仮説 $\mathcal{H}_0$ を考えると、観測度数 f_{ij} は

$$\mu_{ij} = \frac{n_{i\cdot}n_{\cdot j}}{n}$$

に推定されることが分かり、帰無仮説を測る尺度として

$$\chi^2 = \sum_{i=1}^{h}\sum_{j=1}^{k}\frac{(f_{ij}-\mu_{ij})^2}{\mu_{ij}} = \sum_{i=1}^{h}\sum_{j=1}^{k}\frac{(f_{ij}-n_{i\cdot}n_{\cdot j}/n)^2}{n_{i\cdot}n_{\cdot j}/n}$$

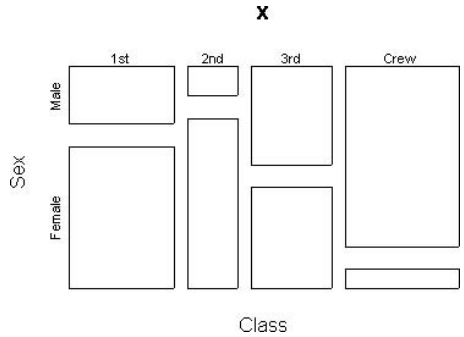
という検定統計量が考えられる。これは前節の考え方を一般化したもので、統計量 χ^2 は自由度 $(k-1)\cdot(h-1)$ のカイ二乗分布に従うことが知られている。

例題 6.3. R のデータ Titanic の成人の男女・船室等級ごとの生存数を考え、生存数に差があるかを考える。有意水準は 5% とする。■

入力と出力結果

```
> x<- Titanic[,,2,2]                          # データ Titanic の成人で生存していた数を x に代入。
> x                                           # x の表示。
      Sex
Class  Male Female
  1st    57    140
  2nd    14     80
  3rd    75     76
  Crew  192     20
> chisq.test(x)                               # x のカイ二乗検定。
        Pearson's Chi-squared test
data:  x                                      # データ名：x。
X-squared = 220.3731, df = 3, p-value < 2.2e-16  # カイ二乗値、自由度、p 値。
```

カイ二乗値 X-squared = 220.3731、自由度 df = 3、p 値は 2.2e-16 以下であるから、帰無仮説：「成人の男女・船室等級間の生存数に差はない」は有意水準 5% で棄却される。このデータのモザイクプロットは右図のようになる。



次に、このデータ Titanic の男女間で生存数が異なるか否かを考える。

入力と出力結果

```
> Sye<- Titanic[,,1,2]+Titanic[,,2,2]         # 子供と成人の生存者数合計。
> Sye<- margin.table(Sye,2)                   # 生存男女の合計。
> Sno<- Titanic[,,1,1]+Titanic[,,2,1]         # 死亡者数合計。
> Sno<- margin.table(Sno,2)                   # 死亡男女合計。
> T<- cbind(Sye, Sno)                         # 生存・死亡、男・女の分割表。
> T                                           # T の表示。
       Sye  Sno
Male   367 1364
Female 344  126
> chisq.test(T)                               # T のカイ二乗検定。
        Pearson's Chi-squared test with Yates'  # Yate の連続補正を伴う、
                        continuity correction   # Pearson のカイ二乗検定。
data:  T                                      # データ名：T
X-squared = 454.4998, df = 1, p-value < 2.2e-16  # カイ二乗値、自由度、p 値。
```

カイ二乗値 `X-squared = 454.4998`、自由度 `df = 1`、p 値は `2.2e-16` 以下であるから、有意水準 5% で帰無仮説：「男女間で生存数に差がない」は棄却される。

例題 6.4. 例題 2.1(17 ページ) のデータ `license` を考え、男女間で差があるかを考える。有意水準は 5% とする。■

入力と出力結果

```
> license
       non bike car
male    18    3  20
female  20   10  15
> chisq.test(license)
        Pearson's Chi-squared test
data:  license
X-squared = 4.4123, df = 2, p-value = 0.1101
```

カイ二乗値 `X-squared = 4.4123`、自由度 `df = 2`、p 値 `0.1101` であるから、有意水準 5% で帰無仮説：「男女差に差がない」は棄却されない。

例題 6.5. 下記の二つの分割表を考える。

データ

	あり	なし	計
男子	4	2	6
女子	1	6	7
計	5	8	13

左のデータを 10 倍したもの

	あり	なし	計
男子	40	20	60
女子	10	60	70
計	50	80	130

上の二つの分割表において「ある」と「なし」は男女間で差があるかを検定せよ。有意水準は 5% とする。■

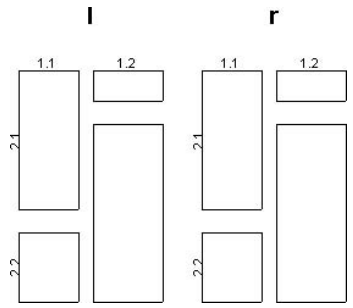
入力と出力結果

```
> data1<- matrix(c(4,1,2,6),nrow=2)                    # データを data1 とする。
> chisq.test(data1)                                    # data1 のカイ二乗検定。
        Pearson's Chi-squared test with Yates' continuity correction
data:  data1                                           # データ名：data1。
X-squared = 1.8591, df = 1, p-value = 0.1727           # カイ二乗値、自由度、p 値。
Warning message:
Chi-squared approximation may be incorrect in: chisq.test(l)
> data10<- 10*data                                     # 10 倍したものを data10 とする。
> chisq.test(data10)                                   # データ data10 のカイ二乗検定。
        Pearson's Chi-squared test with Yates' continuity correction
data:  data10                                          # データ名：data10。
X-squared = 35.272, df = 1, p-value = 2.867e-09        # カイ二乗値、自由度、p 値。
```

元のデータに関しては p 値は `0.1727` となり、有意水準 5% では棄却できないが、10 倍したデータでは p 値は `2.867e-09` となり、有意水準 5% で帰無仮説が棄却される。二つのデータのモザイクプロットを描くと下記のようになり、これからは検定にはあまり役立たないことが分かる。

入力：出力結果は右図

```
> par(mfrow=c(1,2))
> mosaicplot(data1)
> mosaicplot(data10)
> par(mfrow=c(1,1))
```



6.3　正規性の検定

第 4、5 章で考えた検定問題において母集団は正規分布に従っているという仮定をおいてきた。大量生産による工業製品の精度のばらつき、自然界における創造物のような個体の散らばり、測定誤差による散らばり等の理由から正規性の仮定は妥当であると考えられることが多い。節 1.3.7 において正規確率プロットを用いてデータが正規分布に従っているか図式表示することをみたが、ここでは **Shapiro-Wilk 正規性検定**(Shapiro-Wilk test for normality) を用いて母集団が正規分布に従っているか否かの検定問題を扱う。

大きさ n の標本 $X_1, X_2, \cdots, X_n$ が抽出されており、この標本に対して、

帰無仮説 $\mathcal{H}_0$：母集団は正規分布 *versus* 対立仮説 $\mathcal{H}_1$：母集団は正規分布でない

を検定する。標本の順序統計量 $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$ に対して、

$$W = \frac{\left(\sum_{i=1}^{n} a_i X_{(i)}\right)^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2}$$

によって定義される Shapiro-Wilk の正規性検定統計量を考える。ここで、a_i は大きさ n の正規分布からの平均と分散から計算される値で表にされている[*3]。W の値が大きいとき帰無仮説を棄却することになる。

例題 6.6. 平均 5、分散 3^2 の正規母集団からの大きさ 100 の標本について Shapiro-Wilk 正規性検定を考える。また、最小値 2、最大値 4 の一様分布からの大きさ 100 の標本についても正規性の検定を考える。ただし、有意水準は 5% とする。■

入力と出力結果

```
> x<- rnorm(100, mean = 5, sd = 3)     # x に正規乱数を代入。
> y<- runif(100, min = 2, max = 4)     # y に一様乱数を用います。
> par(mfrow=c(1,3))                    # 出力画面を 1*3 に分割。
> boxplot(x,y)                         # x と y の箱ひげ図。図 6.1 の左図参照。
> qqnorm(x); qqline(x)                 # x の正規確率プロット。図 6.1 の真ん中の図参照。
> qqnorm(y); qqline(y)                 # y の正規確率プロット。図 6.1 の右図参照。
> par(mfrow=c(1,1))                    # 出力画面を 1*1 に復元。
> shapiro.test(x)                      # x の Shapiro-Wilk 検定。
        Shapiro-Wilk normality test
data:  x
W = 0.9921, p-value = 0.8308
> shapiro.test(y)                      # y の Shapiro-Wilk 検定。
        Shapiro-Wilk normality test
data:  y
W = 0.9621, p-value = 0.0005658
```

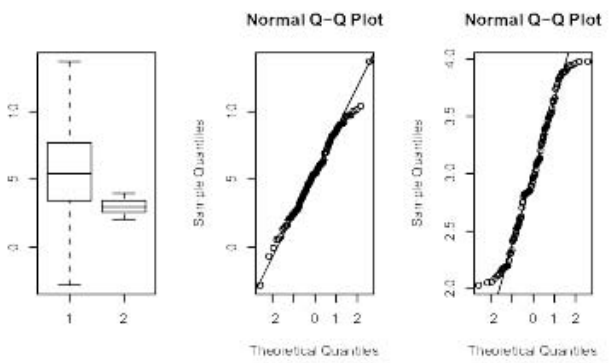


図 6.1　箱ひげ図と正規確率プロット

[*3] Pearson, A. V., and Hartley, H. O. (1972). Biometrica Tables for Statisticians, Vol 2, Cambridge, England, Cambridge University Press.

- 正規確率プロットから x の値はほぼ直線上にあり、y の値は小さな値と大きな値で直線から外れており、正規分布とは異なることを示している。
- データ x に対する Shapiro-Wilk 検定の結果から p 値が `0.8308` であるので、帰無仮説は有意水準 5% で棄却されない。
- データ y に対する Shapiro-Wilk 検定の結果は p 値は `0.0005658` で、帰無仮説は有意水準 5% で棄却される。

6.4 Kolmogorov-Smirnov 検定

Shapiro-Wilk 検定は正規分布に関する検定であったが、**Kolmogorov-Smirnov 検定**(Kolmogorov-Smirnov test) を用いると、他の分布との適合度や二母集団分布の同等性の検定を行うことが出来る。

大きさ n の標本 $X_1, X_2, \cdots, X_n$ に対して、**経験累積分布関数**(empirical cumulative distribution function)

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I[X_i \le x], \qquad -\infty < x < \infty$$

を考える。ここで、関数 $I[A]$ は集合 A が成立するとき、1、成立しないとき、0 をとる関数である。すなわち、x に対して $F_n(x)$ は x 以下のデータの割合を示している。

この関数 $F_n(x)$ は標本数 n が大きいとき、母集団の分布関数 $F(x)$ に近づき[*4]、固定した x に対して

$$\sqrt{n}\Big\{F_n(x) - F(x)\Big\}$$

は、標本数 n が大きいとき近似的に平均ゼロ、分散 $F(x)\{1 - F(x)\}$ の正規分布に従うことが知られている[*5]。

6.4.1 一標本の場合

母集団の分布関数を $F(x)$ とし、想定した分布関数を $F_0(x)$ としたとき

$$\text{帰無仮説 } \mathcal{H}_0:\ F(x) = F_0(x)\ (\forall x) \textit{ versus } \text{対立仮説 } \mathcal{H}_1:\ F(x) \ne F_0(x)\ (\exists x)$$

を検定することを考える。対立仮説は $\mathcal{H}_1:\ F \ne F_0(x)$ という形であるが、片側対立仮説 $\mathcal{H}_1:\ F(x) > F_0(x)$ または、$\mathcal{H}_1:\ F(x) < F_0(x)$ というものも考えることができるが、ここでは扱わない。

経験累積分布関数 $F_n(x)$ に対して、**Kolmogorov-Smirnov 検定統計量**は

$$\begin{aligned} D_n &= x \text{ がすべての範囲を動いたときの } |F_n(x) - F_0(x)| \text{ の最大値} \\ &= \sup_{-\infty<x<\infty} |F_n(x) - F_0(x)| \end{aligned}$$

として定義される。帰無仮説 $\mathcal{H}_0:\ F(x) = F_0(x)$ の下でこの検定統計量の分布は分布関数 $F_0(x)$ に依存しないことを示すことが出来る。また、帰無仮説の下で $F_n(x)$ は $F_0(x)$ に近づくという性質に基づいて考えられた検定統計量で同じような考えに基づき考えられたものとして Cramér-von Mises 統計量などがある。

例題 6.7. Kolmogorov-Smirnov 統計量を用いて、例題 6.6 と同じ検定問題を考える。すなわち、平均 5、分散 3^2 の正規母集団からの大きさ 100 の標本について Kolmogorov-Smirnov 統計量に基づき

$$\text{帰無仮説 } \mathcal{H}_0: \text{母集団は正規分布 } \textit{versus} \text{ 対立仮説 } \mathcal{H}_1: \text{母集団は正規分布でない}$$

を考える。また、最小値 2、最大値 4 の一様分布からの大きさ 100 の標本についても正規性の検定を考える。ただし、有意水準は 5% とする。■

[*4] 確率収束または Glivenko-Cantelli の定理による。

[*5] 中心極限定理による。x の関数として考えたときの確率過程の場合については確率論の本を参考にされたい。

入力と出力結果

```
> ks.test(x, "pnorm", mean=5, sd=3)
        One-sample Kolmogorov-Smirnov test
data:  x
D = 0.0919, p-value = 0.3667
alternative hypothesis: two.sided
> ks.test(y, "pnorm", mean=5, sd=3)
        One-sample Kolmogorov-Smirnov test
data:  y
D = 0.6334, p-value < 2.2e-16
alternative hypothesis: two.sided
```

左図のデータセット x、y は例題 6.6 と同じ値を用いており、同じ検定結果を導いていることが分かる。つまり、正規乱数 x については p 値は 0.3667 で帰無仮説が棄却できなく、一様乱数 y についての p 値は 2.2e-16 であるので正規分布から抽出されているという帰無仮説は棄却されることになる。

Kolmogorov-Smirnov 検定の関数 ks.test の引数はデータ x、分布関数、両側検定か片側検定かを示すもので、関数 ks.test の一般型は

```
ks.test(x, y, ..., alternative = c("two.sided", "less", "greater"), exact = NULL)
```

である。一標本検定を行うとき、引数 y, ..., の部分は分布関数を指定し、

分布	y, ...,
正規分布	pnorm, mean=**, sd=**
一様分布	punif, min=**, max=**
ガンマ分布	pgamma, shape=**, scale=**

のいずれかとなります。

6.4.2　二母集団の比較

第 5 章では、標本 $X_1, X_2, \cdots, X_n$ と標本 $Y_1, Y_2, \cdots, Y_m$ が二つの正規母集団から抽出されていると仮定し、平均の差と分散の比についての検定問題を考えた。ここでは、二つの母集団分布についての正規性の仮定を与えずに、Kolmogorov-Smirnov 検定統計量を用いて、二つの母集団分布が等しいかどうかの検定問題を行う。

標本 X_i が抽出されている母集団分布を $F_X(x)$ によって表し、Y_j が抽出されている母集団分布を $G_Y(x)$ によって表す。ここでの検定問題は

$$\text{帰無仮説 } \mathcal{H}_0 : \ F_X(x) = G_Y(x)\ (\forall x) \ \textit{versus} \ \text{対立仮説 } \mathcal{H}_1 : \ F_X(x) \neq G_Y(x)\ (\exists x)$$

によって与えられる。各標本の経験累積分布関数

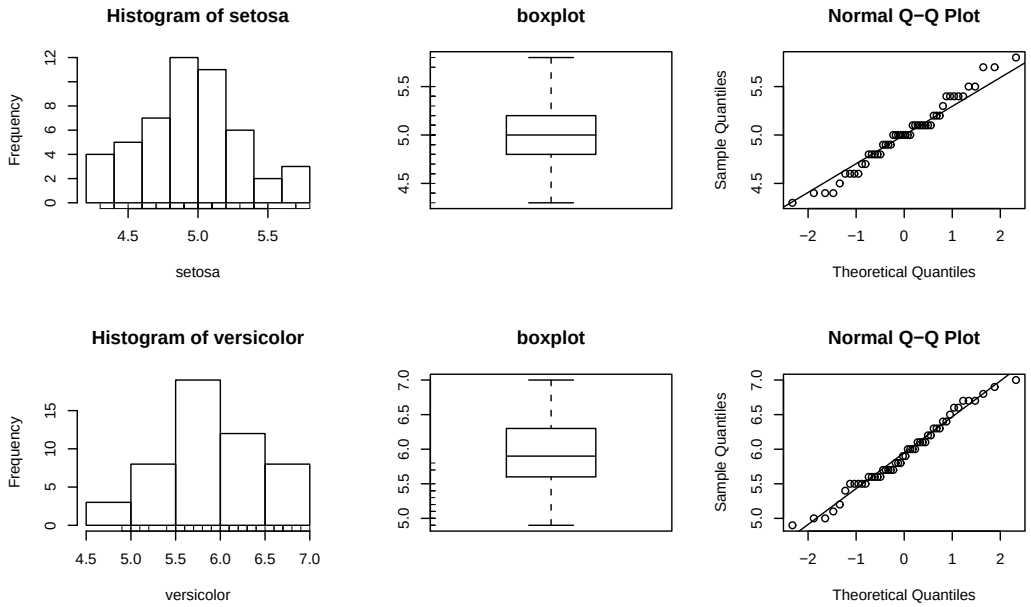
$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I[X_i \le x], \qquad G_m(x) = \frac{1}{m}\sum_{j=1}^{m} I[Y_j \le x], \qquad -\infty < x < \infty$$

を考える。このとき、二つの標本分布の差は

$$\begin{aligned} W_{n,m} &= x \text{ がすべての範囲を動いたときの } |F_n(x) - G_m(x)| \text{ の最大値} \\ &= \sup_{-\infty<x<\infty} |F_n(x) - G_m(x)| \end{aligned}$$

として測ることが出来る。$W_{n,m}$ の値が大きいとき、帰無仮説が成り立っていると考えられなく帰無仮説を棄却することになる。

例題 6.8. R のデータ iris を考え、setosa と versicolor のガク片の長さの分布について Kolmogorov-Smirnov 検定統計量を用いて分布が等しいかどうかの仮説検定を行う。ただし、有意水準は 5% とする。■



上段の三つは `setosa` に対応するヒストグラム、箱ひげ図、正規確率プロットで、下段は `versicolor` に対する三つの図である[*6]。

入力と出力結果

```
> data(iris)                                  # iris の利用。
> setosa<- iris[iris$Species=="setosa",1]     # setosa のガク片長さを setosa とする。
 > versicolor<- iris[iris$Species=="versicolor",1]  # versicolor の定義。
> mean(setosa); var(setosa)                   # setosa の平均、分散。
[1] 5.006
[1] 0.1242490
> mean(versicolor); var(versicolor)           # versicolor の平均、分散。
[1] 5.936
[1] 0.2664327
> shapiro.test(setosa)                        # setosa の Shapiro-Wilk 検定。
        Shapiro-Wilk normality test
data:  setosa
W = 0.9777, p-value = 0.4595
> shapiro.test(versicolor)                    # versicolor の Shapiro-Wilk 検定。
        Shapiro-Wilk normality test
data:  versicolor
W = 0.9778, p-value = 0.4647
```

平均と分散は共に `versicolor` の方が大きいが、Shapiro-Wilk 検定は正規分布を行うと、p 値はそれぞれ、`0.4595` と `0.4647` であり、共に正規分布に従っていると判断して差し支えない。

入力と出力結果

```
> ks.test(setosa, versicolor)                 # Kolmogorov-Smirnov 検定。
        Two-sample Kolmogorov-Smirnov test
data:  setosa and versicolor                  # データ名：setosa と versicolor。
D = 0.78, p-value = 1.230e-13                 # Kolmogorov-Smirnov 検定統計量の値と p 値。
alternative hypothesis: two.sided
```

Kolmogorov-Smirnov 検定統計量の値は 0.78 で p 値は `1.230e-13` とゼロに近いので、二つの分布は異なっているといえる。実際、二つの経験累積分布関数 (`ecdf`) を描くと、図 6.2 から二つの分布はかなりの隔たりを持つことが分かる。

[*6] `par(mfrow=c(3,1)); hist(setosa); rug(setosa); boxplot(setosa); rug(x; side = 2); title("boxplot"); qqnorm(setosa); qqline(setosa);par(mfrow=c(1,1))` として `setosa` の三つの図が描かれている。

入力と出力結果

```
> xlimit<- c(min(setosa,versicolor), max(setosa,versicolor))   # x 軸の調整。
> plot(ecdf(setosa),xlim=xlimit, main="ecdf")                  # setosa の ecdf の描画。
> lines(ecdf(versicolor), lty=2)                               # versicolor の ecdf を線種 2 で描く。
```

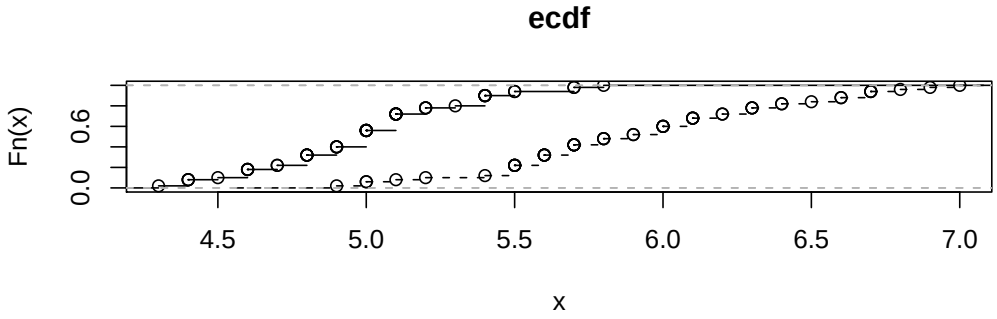


図 6.2　setosa と versicolor の ecdf

問題

問題 6.1.

サイコロを 300 回振って出た目の度数は下記のようになった。有意水準 5% でこのサイコロは公正な (fair) ものかを検定せよ。

目	1	2	3	4	5	6	計
観測度数	54	47	60	54	38	47	300

サイコロを 300 回振った出目の結果は右のようなコマンドで得られます。

入力と出力結果

```
> dice<- 1:6
> r<- sample(dice,300,replace=T)
> table(r)
r
 1  2  3  4  5  6
54 47 60 54 38 47
```

問題 6.2.

ニュースの五つの分類「社会」、「政治」、「経済」、「海外」、「列島トピックス」に関して男女別の意識調査を行ったところ下記の結果を得た。有意水準 5% で男女間に差があるかを検定せよ。

性別	社会	政治	経済	海外	列島トピックス	計
男性	120	34	140	20	50	364
女性	100	60	30	93	30	313
計	220	94	170	113	80	677

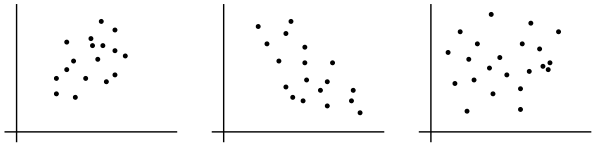
問題 6.3.

データ iris の種 setosa のガク片の長さは正規分布に従うかを有意水準 5% で検定せよ。

第 7 章　相関分析と単回帰分析

7.1　散布図

n 個の個体に対して、対の観測を伴うデータ $(X_1, Y_1), (X_2, Y_2), \cdots, (X_n, Y_n)$ が与えられているとき、$x-y$ 平面上にこれらのデータを打点をした図を**散布図**(scatter diagram) という。



(1)　(2)　(3)

図 7.1　散布図

図 7.1 (1) は x が増加すると y も増加している、図 7.1 (2) はその逆で x が増加すると y は減少しているようにみられ、図 7.1 (3) は x と y の間には関係がないようにみられる。これらの性質を数値化したものの中で代表的なものは Pearson による**相関係数**(correlation coefficient)

$$r = \frac{s_{XY}}{\sqrt{s_{XX} s_{YY}}} \tag{7.1}$$

である。ここで、

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad \overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i, \quad s_{XY} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{n-1}, \quad s_{XX} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}, \quad s_{YY} = \frac{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}{n-1}$$

である。

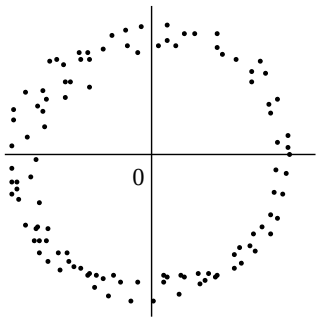
相関係数 r の値についてはシュワルツ (Schwarz) の不等式より、

$$-1 \leq r \leq 1$$

が成り立ち、

- $r > 0$ のとき、正の相関を持つ。
- $r < 0$ のとき、負の相関をもつ。
- $|r|$ の値が 1 に近いとき、x と y の間に直線的な関係がある。

という。r の値がゼロに近くても関連のある場合があり、ピアソンの相関係数は直線的な相関を測る尺度になっている。例えば、下図の散布図にあるようなデータは相関係数がゼロに近いが (X_i, Y_i) の間には関連がある。

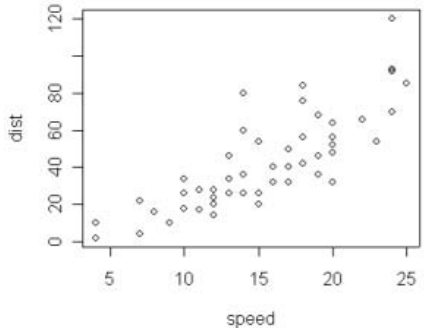


相関係数は 0 に近いが、相関がある散布図

例題 7.1. 3 ページと 28 ページで扱ったデータ cars を考える。cars の相関係数行列を求める。■

入力と出力結果

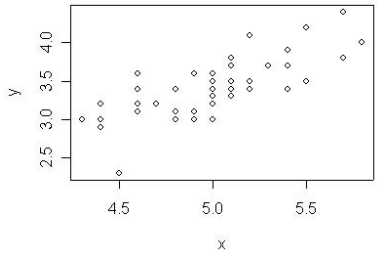
```
> data(cars)
> plot(cars)                    # データの散布図。
> cor(cars)                     # 相関係数行列の計算。
          speed     dist
speed 1.0000000 0.8068949
dist  0.8068949 1.0000000
```



例題 7.2. 11 ページで扱ったデータ iris の種 setos ガク片の長さと幅を考え、散布図と相関係数を求める。■

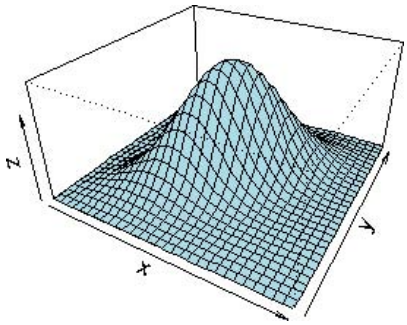
入力と出力結果

```
> data(iris)
> setos<- iris[iris$Species=="setosa",]   # setosa だけ抜き出す。
> x<- iris[setos,1]                       # setosa のガク片の長さ。
> y<- iris[setos,2]                       # setosa のガク片の幅。
> plot(x,y)                               # データの散布図。
> cor(x,y)                                # 相関係数。
[1] 0.7425467
```



n 個の組のデータ $(X_1, Y_1), (X_2, Y_2), \cdots, (X_n, Y_n)$ が二変量正規分布母集団から抽出された標本と考えることが多い。この場合の母集団分布関数は確率密度関数

$$f(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho_{XY}^2}} \exp\left[-\frac{1}{2(1-\rho_{XY}^2)}\left\{\frac{(x-\mu_X)^2}{\sigma_X^2} - 2\rho_{XY}\frac{x-\mu_X}{\sigma_X}\cdot\frac{y-\mu_Y}{\sigma_Y} + \frac{(y-\mu_Y)^2}{\sigma_Y^2}\right\}\right]$$



によって与えられる。ここで、μ_X、σ_X^2 は各々 X の平均と分散で、μ_Y、σ_Y^2 は各々 Y の平均と分散で、X と Y の相関係数は

$$\rho_{XY} = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}} = \frac{\mathrm{Cov}[X,Y]}{\sigma_X\sigma_Y}$$

によって定義される。二変量正規分布の特徴として、

$$\rho_{XY} = 0 \qquad \Leftrightarrow \qquad X \text{ と } Y \text{ は独立}$$

というものがあり、標本から計算された相関係数がゼロに近いとき変量間に関連がないというのは正規分布を仮定したときに成り立つことである。右上の図は下記のコマンドにより描かれた二変量正規密度関数 $(\mu_X = \mu_Y = 0, \sigma_X = \sigma_X = 1, \rho_{XY} = 0.55)$ である。

入力

```
> x<- seq(-3,3,length=30); y<- seq(-3,3,length=30)
> f <- function(x,y) {1/(2*pi*sqrt(1-0.55^2))*exp(-1/(2*(1-0.55^2))*(x^2-2*0.55*x*y+y^2))
> z <- outer(x, y, f)
> persp(x, y, z, theta = 30, phi = 30, expand = 0.5, col = "lightblue")
```

三変数以上の正規分布についてはベクトルと行列の記号・演算を用いて定義した方が便利である。確率ベクトル $\mathbf{X} = (X_1, X_2, \cdots, X_n)^t$ が n 次元正規分布 $\mathcal{N}(\mu, \boldsymbol{\Sigma})$ に従うとは、その確率密度関数が

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{n/2}\{\det(\boldsymbol{\Sigma})\}^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\mu)^t\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\mu)\right\}$$

であたえられるものをいう。ここで、^{t} はベクトルの転置を、$^{-1}$ は行列の逆行列を、det は行列の行列式を意味し、

$$\mathbf{x} = (x_1, x_2, \cdots, x_n)^t, \quad \mu = (\mu_1, \mu_2, \cdots, \mu_n)^t,$$
$$\mathbf{\Sigma} = \begin{pmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 & \rho_{13}\sigma_1\sigma_3 & \cdots & \rho_{12}\sigma_1\sigma_n \\ \rho_{21}\sigma_2\sigma_1 & \sigma_2^2 & \rho_{23}\sigma_2\sigma_3 & \cdots & \rho_{2n}\sigma_2\sigma_n \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho_{n1}\sigma_n\sigma_1 & \rho_{n2}\sigma_n\sigma_2 & \rho_{n3}\sigma_n\sigma_3 & \cdots & \rho_{nn}\sigma_n^2 \end{pmatrix},$$

μ_i、σ_i^2 はそれぞれ変数 X_i の平均と分散であり、ρ_{ij} は変数 X_i と X_j の間の相関係数である。

7.2　相関係数の検定

母集団分布は二変量正規分布

$$\mathcal{N}\left(\begin{pmatrix}\mu_1 \\ \mu_2\end{pmatrix}, \begin{pmatrix} \sigma_X^2 & \rho_{XY}\sigma_X\sigma_Y \\ \rho_{XY}\sigma_X\sigma_Y & \sigma_Y^2 \end{pmatrix}\right)$$

に従うと仮定し、この母集団から抽出された大きさ n の標本 $(X_i, Y_i), i = 1, 2, \cdots, n$ に基づき、相関係数 ρ_{XY} の仮説検定を考える。

帰無仮説の形

$$\text{帰無仮説 } \mathcal{H}_0 : \ \rho_{XY} = 0$$
$$\text{帰無仮説 } \mathcal{H}_0 : \rho_{XY} = \rho_0$$

によって検定方式が異なっている。ここで、ρ_0 はあらかじめ定められた値である。

7.2.1　帰無仮説 $\mathcal{H}_0 : \rho_{XY} = 0$ の場合

帰無仮説 $\mathcal{H}_0 : \rho_{XY} = 0$ の下で、標本 $(X_i, Y_i), i = 1, 2, \cdots, n$ は二つの独立な正規分布からの標本と考えることができ、式 (7.1) によって定義される標本相関係数を変換して、検定統計量

$$t_0 = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

を考える。この検定統計量 t_0 は自由度 $n-2$ の t 分布に従う[*1]ことが知られており、下記のような棄却領域が作られる。これは無相関の検定とも呼ばれている。

$$\text{帰無仮説 } \mathcal{H}_0\text{: } \rho_{XY} = 0 \textit{ versus } \text{対立仮説 } \mathcal{H}_1\text{: } \rho_{XY} \neq 0 \quad \begin{cases} |t| > t_{n-2}(\alpha/2) \text{ ならば帰無仮説を棄却する。} \\ |t| \leq t_{n-2}(\alpha/2) \text{ ならば帰無仮説は棄却できない。} \end{cases}$$

$$\text{帰無仮説 } \mathcal{H}_0\text{: } \rho_{XY} = 0 \textit{ versus } \text{対立仮説 } \mathcal{H}_1\text{: } \rho_{XY} > 0 \quad \begin{cases} t > t_{n-2}(\alpha) \text{ ならば帰無仮説を棄却する。} \\ t \leq t_{n-2}(\alpha) \text{ ならば帰無仮説は棄却できない。} \end{cases}$$

$$\text{帰無仮説 } \mathcal{H}_0\text{: } \rho_{XY} = 0 \textit{ versus } \text{対立仮説 } \mathcal{H}_1\text{: } \rho_{XY} < 0 \quad \begin{cases} t < -t_{n-2}(\alpha) \text{ ならば帰無仮説を棄却する。} \\ t \geq -t_{n-2}(\alpha) \text{ ならば帰無仮説は棄却できない。} \end{cases}$$

例題 7.3. 例題 7.1(64 ページ) で扱ったデータ `cars` を考え、`cars` の `speed` と `dist` の相関係数 ρ_{XY} とするとき検定問題

$$\text{帰無仮説 } \mathcal{H}_0\text{: } \rho_{XY} = 0 \textit{ versus } \text{対立仮説 } \mathcal{H}_1\text{: } \rho_{XY} \neq 0$$

を考える。ただし、有意水準を 5% とする。■

[*1] t 分布については 35 ページ参照。

入力と出力結果

```
> data(cars)
> cor.test(cars$speed, cars$dist, alternative="two.sided")   # speed と dist の相関係数検定。
        Pearson's product-moment correlation                 # Pearson の積率相関係数検定。
data:  cars$speed and cars$dist                              # データ名：cars。
t = 9.464, df = 48, p-value = 1.49e-12                       # t-値、自由度、p 値。
alternative hypothesis: true correlation                     # 対立仮説：「相関係数は 0 でない」。
           is not equal to 0
95 percent confidence interval:                              # 95% 信頼区間。
 0.55816422 0.8862036                                        # (0.55816422, 0.8862036)
sample estimates:                                            # 標本推定値。
      cor                                                    # 相関係数。
0.8068949
```

p 値は 1.49e-12 となっており、有意水準 5% で帰無仮説：「相関係数はゼロである」が棄却される。上の計算の検算は右のようになる。

入力と出力結果

```
> r<- cor(cars$speed, cars$dist)
> n<- length(cars$speed)
> r*sqrt(n-2)/sqrt(1-r^2)
0.8068949
```

例題 7.4. 例題 7.2(64 ページ) で扱ったデータ iris の種 setosa のガク片の長さと幅の相関係数 ρ_{XY} についての検定問題

$$\text{帰無仮説 } \mathcal{H}_0\colon \rho_{XY} = 0 \textit{ versus } \text{対立仮説 } \mathcal{H}_1\colon \rho_{XY} \neq 0$$

を考える。ただし、有意水準を 5% とする。■

入力と出力結果

```
> data(iris)
> x<- iris[iris$Species=="setosa",1]                         # ガク片の長さを x に代入。
> y<- iris[iris$Species=="setosa",2]                         # ガク片の幅を y に代入。
> cor.test(x,y,alternative="two.sided")                      # x と y の積率相関検定。
        Pearson's product-moment correlation                 # Pearson の積率相関検定。
data:  x and y                                               # データ名：x と y。
t = 7.6807, df = 48, p-value = 6.71e-10                      # t-値、自由度、p 値。
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:                              # 95% 信頼区間。
 0.5851391 0.8460314                                         # (0.5851391, 0.8460314)
sample estimates:                                            # 標本推定値。
      cor                                                    # 相関係数。
0.7425467
```

p 値は 6.71e-10 となっており、有意水準 5% で帰無仮説：「相関係数はゼロである」が棄却される。

同じようにして、データ cars の相関係数についての片側検定問題

$$\text{帰無仮説 } \mathcal{H}_0\colon \rho_{XY} = 0 \textit{ versus } \text{対立仮説 } \mathcal{H}_1\colon \rho_{XY} > 0$$

に対しては、コマンド

入力

```
> data(cars)
> cor.test(cars$speed, cars$dist, alternative="greater")   # speed と dist の積率相関係数検定
```

を用い、片側検定問題

$$\text{帰無仮説 } \mathcal{H}_0\colon \rho_{XY} = 0 \textit{ versus } \text{対立仮説 } \mathcal{H}_1\colon \rho_{XY} < 0$$

に対しては、コマンド

入力

```
> data(cars)
> cor.test(cars$speed, cars$dist, alternative="less")   # speed と dist の積率相関係数検定
```

を用いることになります。関数 cor.test は今まで検定問題に用いた関数 t.test、var.test と同様に対立仮説の形を alternative="tow.sided"、"greater"、"less"によって指定します。

7.2.2　帰無仮説 $\mathcal{H}_0 : \rho_{XY} = \rho_0$ の場合

帰無仮説 $\mathcal{H}_0 : \rho_{XY} = \rho_0$ についての仮説検定を行うために、式 (7.1) によって定義される標本相関係数 r を変換した統計量

$$z = \tanh^{-1} r \fallingdotseq \frac{1}{2} \log \frac{1+r}{1-r} \tag{7.2}$$

を用いる。これはフィッシャー (Fisher) の z 変換と呼ばれおり、標本数 n が大きいとき、正規分布

$$\mathcal{N}\left(\zeta, \frac{1}{n-3}\right)$$

に従うことを示すことが出来、標本数 n が 10 以上であれば正規分布で近似しても差し支えないことが知られている。ここで、

$$\zeta = \frac{1}{2} \log \frac{1+\rho_0}{1-\rho_0}$$

である。従って、帰無仮説 $\mathcal{H}_0 : \rho_{XY} = \rho_0$ の下で、z を標準化した量

$$u_0 = \sqrt{n-3}\left(z - \frac{1}{2} \log \frac{1+\rho_0}{1-\rho_0}\right)$$

は近似的に標準正規分布 $\mathcal{N}(0, 1^2)$ に従う。

以上のことにより、有意水準を α とするとき下記のような棄却領域が作られる。

$$\text{帰無仮説 } \mathcal{H}_0\colon \rho_{XY} = \rho_0 \textit{ versus } \text{対立仮説 } \mathcal{H}_1\colon \rho_{XY} \neq \rho_0 \quad \begin{cases} |u_0| > z_{\alpha/2} \text{ ならば帰無仮説を棄却する。} \\ |u_0| \leq z_{\alpha/2} \text{ ならば帰無仮説は棄却できない。} \end{cases}$$

$$\text{帰無仮説 } \mathcal{H}_0\colon \rho_{XY} = \rho_0 \textit{ versus } \text{対立仮説 } \mathcal{H}_1\colon \rho_{XY} > \rho_0 \quad \begin{cases} u_0 > z_{\alpha} \text{ ならば帰無仮説を棄却する。} \\ u_0 \leq z_{\alpha} \text{ ならば帰無仮説は棄却できない。} \end{cases}$$

$$\text{帰無仮説 } \mathcal{H}_0\colon \rho_{XY} = \rho_0 \textit{ versus } \text{対立仮説 } \mathcal{H}_1\colon \rho_{XY} < \rho_0 \quad \begin{cases} u_0 < -z_{\alpha} \text{ ならば帰無仮説を棄却する。} \\ u_0 \geq -z_{\alpha} \text{ ならば帰無仮説は棄却できない。} \end{cases}$$

例題 7.5. 例題 7.1(64 ページ) で扱ったデータ cars を考え、speed と dist の相関係数 ρ_{XY} とするとき検定問題

$$\text{帰無仮説 } \mathcal{H}_0\colon \rho_{XY} = 0.8 \textit{ versus } \text{対立仮説 } \mathcal{H}_1\colon \rho_{XY} \neq 0.8$$

を考える。ただし、有意水準を 5% とする。■

入力と出力結果

```
> r0<- 0.8                                  # 対立仮説の相関係数の値。
> data(cars)
> x<- cars$speed                            # speed を x とする。
> y<- cars$dist                             # dist を y とする。
> r<- cor(x,y)                              # x と y の相関係数を r とする。
> n<- length(x)                             # データ数を n とする。
> z<- 0.5*log((1+r)/(1-r))                  # r の z 変換。
> sqrt(n-3)*(z-0.5*log((1+r0)/(1-r0)))      # z を標準化。
[1] 0.1333629
```

u_0 の値は 0.1333629 となっており、有意水準 5% の棄却限界値 $z_{0.25} = 1.96$ より小さいので、帰無仮説：「相関係数は 0.8 である」は棄却することは出来ない。28 ページにあるように speed と dist の相関係数は 0.8068949 である。

同様に、片側検定問題

$$\text{帰無仮説 } \mathcal{H}_0\colon \rho_{XY} = \rho_0 \textit{ versus } \text{対立仮説 } \mathcal{H}_1\colon \rho_{XY} > \rho_0$$

と

$$\text{帰無仮説 } \mathcal{H}_0\colon \rho_{XY} = \rho_0 \textit{ versus } \text{対立仮説 } \mathcal{H}_1\colon \rho_{XY} < \rho_0$$

に対しては、上で用いたのと同じコマンドを用いて検定を行うことができる。

7.3　単回帰分析

両親の身長が子供の身長にどのような影響を及ぼすか、データ `cars` の `speed` と `dist` の間に関連があるか、データ `sleep` の睡眠薬の違いが睡眠時間の差を生じさせるか等、説明したい変量を原因と思われる変量を用いて記述したいことがある。このとき、原因と考えられる量 (変数) を**説明変数**(explanatory variable) または**独立変数**(independent variable) といい、説明したい量 (変数) を**目的変数**(criterion variable) または**従属変数**(dependent variable) という。

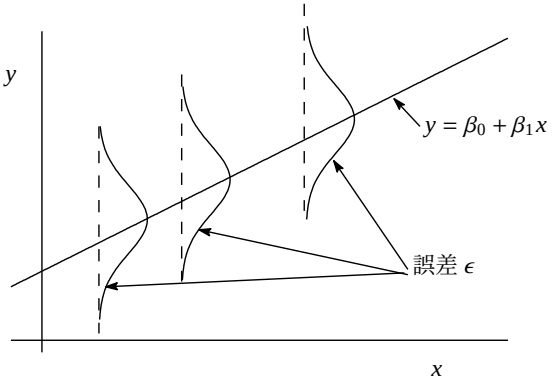
ここでは、連続尺度データ x_i に対して連続尺度データ y_i が観測され、**回帰係数**(regression coefficient) β_0 と β_1 を用いて、

$$y_i = \beta_0 + \beta_1(x_i - \overline{x}) + \epsilon_i \qquad i = 1, 2, \cdots, n \tag{7.3}$$

と表現される**線形単回帰モデル**(simple regression model) を考える。ここで、ϵ_i は観測誤差と呼ばれ、x_i を与える実験に伴い発生する誤差を表し、説明変数だけでは表すことの出来ない偶然変動と考える。二つ以上の説明変数の線形式を用いて、目的変数を記述するモデルを**線形重回帰モデル**(multiple regression model) といい、説明変数の二次式および対数式を含むモデルを**非線形回帰モデル**(non-linear regression model) という。

- x_i が名義尺度または順序尺度データ、y_i が連続尺度データのとき、一元配置モデル、
- x_i が連続尺度データ、y_i が名義尺度または順序尺度データのとき、ロジスティックモデル、
- x_i が名義尺度または順序尺度データ、y_i が名義尺度または順序尺度データのとき、分割表モデル

と呼ぶ。ここでは、x_i と y_i が共に連続尺度データである線形単回帰モデルのみを扱う。



式 (7.3) によって与えられる線形単回帰モデルを考える際、観測誤差 ϵ_i について

- 平均 0、分散 σ^2 の正規分布に従う。
- $i \neq j$ に対して、ϵ_i と ϵ_j は独立である。

を仮定する。これは観測値 y_i は、誤差 ϵ_i を伴い $\beta_0 + \beta_1 x_i$ を平均として、互いに関連なく観測されているということを意味している。

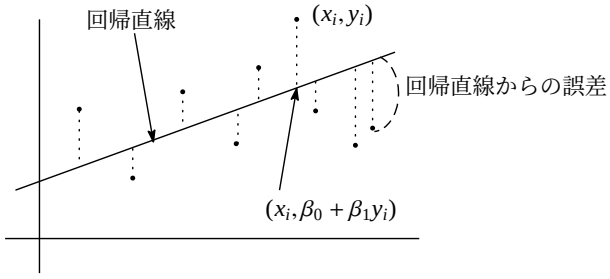
誤差が正規分布に従っているという仮定の下で、この線形単回帰モデル (式 (7.3)) の回帰係数 β_0 と β_1 を推定するために、誤差の総和

$$L = \sum_{i=1}^{n} \left\{ y_i - \beta_0 - \beta_1(x_i - \overline{x}) \right\}^2$$

を最小とする最小二乗法を用いる。この L を最小にする解は

$$\hat{\beta}_0 = \overline{y}, \qquad \hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} = \frac{s_{xy}}{s_{xx}}$$

で与えられる[*2]。ここで



$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad \overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i, \quad s_{xy} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{n-1}, \quad s_{xx} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}$$

[*2] L を β_0 と β_1 に関して偏微分したものをゼロとし、方程式を解いたものである。この方程式は正規方程式と呼ばれる。

である。

例題 7.6. 10 人の家計調査をしたところ、収入と支出が下記のようであった。

収入	70.2	60.1	57.5	54.9	62.4	61.1	55.7	56.4	58.5	54
支出	38.3	32.6	32.7	34.9	35.1	36.6	32.3	31.4	34.9	31.8

支出を収入で記述することを考える。■

入力：出力結果は図 7.2

```
> x<- c(70.2, 60.1, 57.5, 54.9, 62.4, 61.1, 55.7, 56.4, 58.5, 54)    # 収入を x とする。
> y<- c(38.3, 32.6, 32.7, 34.9, 35.1, 36.6, 32.3, 31.4, 34.9,31.8)   # 支出を y とする。
 > plot(x,y)                                                         # x に対する y の散布図描画。
> reg1<- lm(y~x)                                                     # x に対する y の線形回帰分析。
> abline(reg1)                                                       # x に対する y の回帰直線。
```

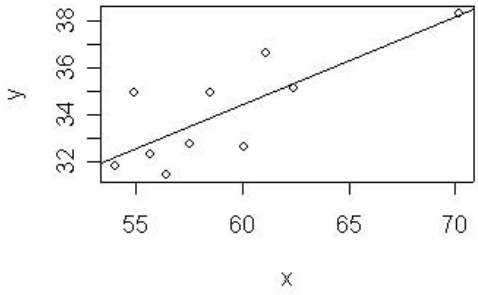


図 7.2　収入と支出の散布図と回帰直線

線形回帰に対する詳細な数値結果は関数 summary によって求まります。

入力と出力結果

```
> summary(re1)                                                       # reg1 の要約。
Call:
lm(formula = y ~ x)                                                  # モデル。
Residuals:                                                           # 残差。
    Min      1Q  Median      3Q     Max                             # 残差の要約値。
-1.8438 -0.55962 -0.2789  0.8077  2.4128
Coefficients:                                                        # 回帰係数。
            Estimate Std. Error t value Pr(>t)|
(Intercept)  11.8303     6.0805   1.946  0.08758 .                  # 切片の推定値、標準誤差、t-値、確率。
x             0.3763     0.1026   3.667  0.00634 **                 # 係数の推定値、標準誤差、t-値、確率。
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1      # 上の 2 行の右端*の意味。
Residual standard error: 1.468 on 8 degrees of freedom               # 残差標準誤差、自由度。
Multiple R-Squared: 0.55269,     Adjusted R-squared: 0.5803          # 決定係数、調整済みの決定係数。
F-statistic: 13.44 on 1 and 8 DF,  p-value: 0.006341                 # F-統計量、p 値。
```

上の数値は下記のような意味を持っている。

- 切片と傾きの推定値はそれぞれ 11.8303、0.3763 であるので、回帰式

$$y = 11.8303 + 0.3763 \times x$$

  を得る。
- 誤差の標準誤差は 1.468 である。

- 決定係数は 0.55269 で、これは支出 y のバラツキの 55.3% が収入で説明できることを意味している。この値が 1 に近いほど、モデルは適切なものになる。なお、決定係数は相関係数の二乗である。
- F 統計量は回帰による要因の不偏分散を回帰からの残差の不偏分散で割ったもので、この値は自由度 $(1, n-2)$ の F 分布に従うことが知られており、p 値が `0.006341` であるので、回帰モデルは有効であることが分かる。

例題 7.7. データ `cars` を考える。停止にようする距離 `dist` を車の速度 `speed` で表すことを考える。■

入力：出力結果は図 7.3

```
> data(cars)
> plot(dist~speed, data=cars)        # speed に対する dist の散布図描画。
> reg1<-lm(dist~speed, data=cars)    # dist を speed で表したときの線形回帰の結果を reg1 に代入。
> abline(reg1)                       # reg1 の傾きと切片を持つ直線を描画。
```

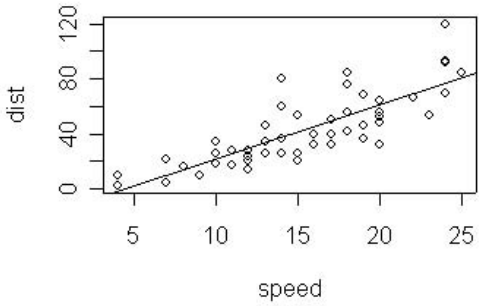


図 7.3　`cars` の `speed` と `dist` の散布図と回帰直線

線形回帰に対する詳細な数値結果 `summary(reg1)` は次のようになります。

入力と出力結果

```
> summary(re1)                                                    # reg1 の要約。
Call:
lm(formula = dist ~ speed, data = cars)                           # モデル。
Residuals:                                                        # 各観測値対する残差の要約値。
    Min      1Q  Median      3Q     Max
-29.069  -9.525  -2.272   9.215  43.201
Coefficients:                                                     # 回帰係数。
            Estimate Std. Error t value Pr(>t)|
(Intercept) -17.5791     6.7584  -2.601   0.0123 *               # 切片の推定値、標準誤差、t-値、確率。
speed         3.9324     0.4155   9.464 1.49e-12 ***             # 係数の推定値、標準誤差、t-値、確率。
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1   # 上の 2 行の右端*の意味。
Residual standard error: 15.38 on 48 degrees of freedom           # 残差標準誤差、自由度。
Multiple R-Squared: 0.55511,     Adjusted R-squared: 0.55438      # 決定係数、調整済みの決定係数。
F-statistic: 89.57 on 1 and 48 DF,  p-value: 1.490e-12            # F-統計量、p 値。
```

上の数値は下記のような意味を持っている。

- 切片と傾きの推定値はそれぞれ-17.5791、3.9324 であるので、回帰式

$$y = -17.5791 + 3.9324 \times x$$

  を得る。
- 誤差の標準誤差は 15.38 である。
- 決定係数は 0.55511 で、これは支出 y のバラツキの 55.5% が、収入で説明できることを意味している。この値が 1 に近いほど、モデルは適切なものになる。

- p 値が `1.490e-12` であるので、回帰モデルは有効であることが分かる。

説明変数が二つ以上になる重回帰分析では解析に必要ないと思われる変数や重複している変数があり、上の解析の数値で現れた回帰係数の t-値が重要な意味を持つことになる。

問題

問題 7.1.

データ `faithful` を考える。これは Yellowstone National Park の Old Faithful 間欠泉[*3]の噴出時間 (`eruptions`) と噴出間隔 (`waiting`) の 272 個の観測値である。

(i) 噴出間隔を噴出時間で説明することを考える。
- 散布図と回帰直線を描画せよ。
- 回帰直線を求めよ。

(ii) 噴出時間を噴出間隔で説明することを考える。
- 散布図と回帰直線を描画せよ。
- 回帰直線を求めよ。

[*3] 熱湯や水蒸気が周期的に断続して噴出するもの。

第 8 章　品質管理

今までの章でヒストグラム、パレート図、散布図を見たが、これは品質管理の七つ道具と呼ばれているもののうちの三つであり、この章では残りの四つの手法 (特性要因図、管理図、チェックーシート、層別) について考える。すでに見たように、品質管理のパッケージ qcc があるので、これを組み込んでデータ処理を行う。

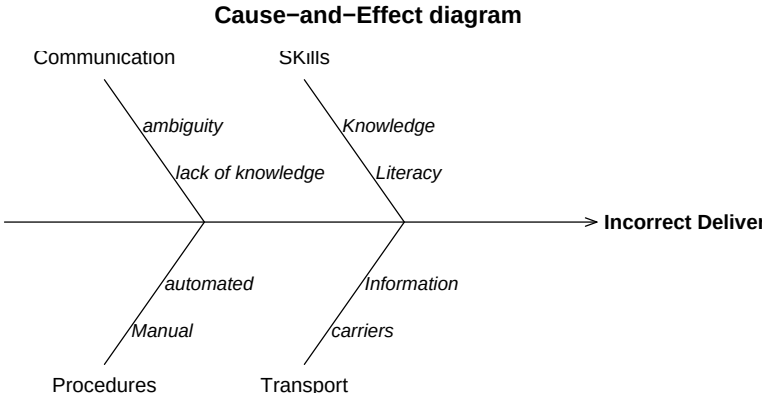
8.1　特性要因図

特性要因図(cause and effect diagram) とは、特性と要因の関係を図式するものであり、結果と原因や目的と手段等の関係を図示するために用いられ、魚の骨グラフ (Fishbone Diagram) とも呼ばれる。

例題 8.1. 配達の遅れについて考える。要因の大分類としては、「Communication」、「Skills」、「Procedures」、「Transport」を考え、各大分類の下に細かな要因を付ける。■

入力

```
> library(qcc)                                          # ライブラリ qcc の組み込み。
> cause.and.effect(cause=list(
>    Communication=c("ambiguity", "lack of knowledge"),
>    Skills=c("Knowledge", "Literacy"),
>    Procedures=c("Manual", "automated"),
>    Transport=c("carriers", "Information")),
>    effect="Incorrect Deliver")
```



8.2　管理図

製品の工程管理を行っていく際、製品の品質・精度が一定でなく、ばらつくことがある。このばらつきを偶然原因によるものか異常原因によるものかに区別し、異常原因を取り除くことが重要となる。このとき利用される図がシュハート (Shewhart) により提案された**管理図**(control chart) であり、**中心線**(CL：Center Line)、**上部管理限界線**(UCL：Upper Control Limit) と**下部管理限界線**(LCL：Lower Control Limit) の三本を配した図にデータを打点したものである。

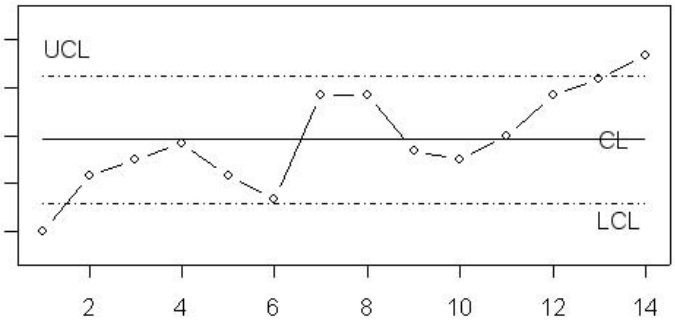


図 8.1　管理図の一例

打点されるデータは通常ひとまとまりの集まり (群、ロット) の代表値を用いたり、群内のすべてのデータを用いたりする。この推移グラフから

- 打点した点が中心線の上方または下方に偏る。
- 打点した点に周期性がある。
- 連続して下降または上昇する連続した五つ以上の点がある。
- 上部または下部管理限界線を超えた点がある。

などの項目に注意し、製品の製造工程を管理・維持していくことになる。

ライブラリ qcc には標本から群に対応したデータセットを作成するための関数 qcc.groups があり、群の構造を持つデータを構成します。

例題 8.2. 14 日間にわたり、朝、昼、夜の体重測定をした結果が下記のものとなった。品質管理として扱うデータの形式を作ることを考える。■

日	朝	昼	夜
1	52.1	52.1	52.4
2	52.3	52.2	52.8
3	52.5	52.4	52.6
4	52.6	52.6	52.5
5	52.7	52.3	52.3
6	52.3	52.3	52.4
7	52.6	52.7	53.0
8	52.8	52.7	52.8
9	52.6	52.5	52.5
10	52.4	52.5	52.6
11	52.5	52.6	52.7
12	52.8	52.7	52.8
13	52.7	52.8	53.0
14	52.8	52.9	53.1

入力と出力結果

```
> library(qcc)                                  # ライブラリ qcc の利用。
> weight<- matrix(c(52.1,52.1,52.4,             # 体重データを weight に代入。
+ 52.3,52.2,52.8,
                  ⋮                             <<途中略>>
+ 52.7,52.8,53.0,
+ 52.8,52.9,53.1),nrow=14)                      # 14 行のデータ。
> sample<- rep(1:14, each=3)                    # 14 行のデータに番号を付ける。
 > d<- qcc.groups(weight, sample)               # 関数 qcc.groups を用いて変換。
> colnames(d)<- c("morning", "day","night")     # 列名を付値。
> d                                             # データ d の確認。
```

8.2.1　連続型の場合

観測されているデータが連続型で、群間に起きている変動を調べたい場合と群内の変動を調べたい場合により、管理図が異なってくる。抽出されるデータを $x_{ij}, i = 1, 2, \cdots, k, j = 1, 2, \cdots, n$ によって表し、i で第 i 群を表し、各群には n 個のデータがあるものとする。通常、群内の標本数 n は 4、5、6 にとられることが多い。

$\overline{x}$ 管理図

郡内平均 $\overline{x}_{i\cdot}$ に関する管理図で、各群の平均値

$$\overline{x}_{i\cdot} = \frac{\sum_{j=1}^{n} x_{ij}}{n}, \qquad i = 1, 2, \cdots, k$$

を打点し、中心線 (CL)、上部管理限界線 (UCL)、下部管理限界線 (LCL) をそれぞれ

$$\mathrm{CL} = \frac{\sum_{i=1}^{k} \overline{x}_{i\cdot}}{k} = \overline{\overline{x}}, \quad \mathrm{UCL} = \overline{\overline{x}} + A_2\overline{R}, \quad \mathrm{LCL} = \overline{\overline{x}} - A_2\overline{R}$$

によって与える。ここで、係数 A_2 は右の表で与えられ、

$$\overline{R} = \frac{\sum_{i=1}^{k} R_i}{k}, \qquad R_i = \max_{1\le j\le n} x_{ij} - \min_{1\le j\le n} x_{ij} = \text{第 } i \text{ 群の範囲}$$

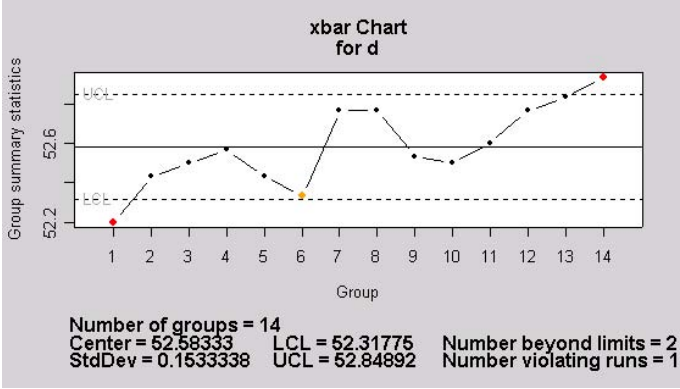
である。

群の大きさ (k)	A_2
2	1.88
3	1.023
4	0.729
5	0.577
6	0.483
7	0.419
8	0.373
9	0.337
10	0.308

この節の始めで扱った体重のデータについての $\overline{x}$ 管理図を描くと下記のようになる。

入力と出力結果

```
> qcc(d, type="xbar")
Call:
qcc(data = d, type = "xbar")
xbar chart for d
Summary of group statistics:
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  52.20   52.45   52.55   52.58   52.77   52.93
Group sample size:  3
Number of groups:  14
Center of group statistics:  52.58333
Standard deviation:  0.1533338
Control limits:
      LCL      UCL
 52.31775 52.84892
```



最後の期間において平均についての上昇傾向 (連という) がみられ、第 1 日と 14 日目に管理限界を超えている。

S 管理図

郡内分散 s_i に関する管理図で、各群内の標本分散

$$s_i = \frac{\sum_{j=1}^{n}(x_{ij} - \overline{x}_{i\cdot})^2}{n-1}, \qquad i = 1, 2, \cdots, k$$

を打点し、中心線 (CL)、上部管理限界線 (UCL)、下部管理限界線 (LCL) をそれぞれ

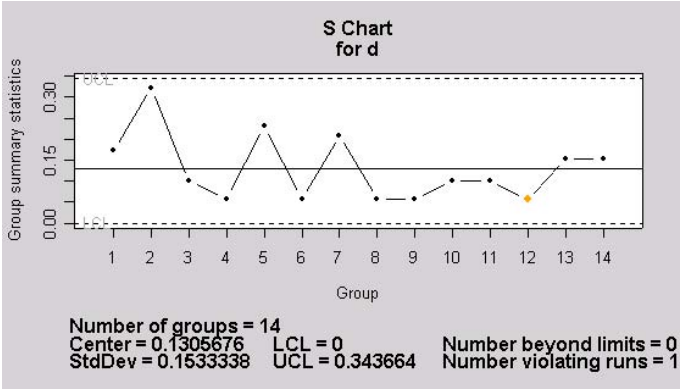
$$\mathrm{CL} = \frac{\sum_{i=1}^{k} s_i}{k} = \overline{s}, \quad \mathrm{UCL} = B_4\overline{s}, \quad \mathrm{LCL} = B_3\overline{s}$$

によって与える。ここで、係数 B_3、B_4 は右の表で与えらる。

群の大きさ (k)	B_3	B_4
2	0	3.267
3	0	2.575
4	0	2.282
5	0	2.115
6	0.030	2.004
7	0.118	1.924
8	0.185	1.864
9	0.239	1.816
10	0.284	1.777

入力と出力結果

```
> qcc(d, type="S")
Call:
qcc(data = d, type = "S")
S chart for d
Summary of group statistics:
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.05774 0.05774 0.10000 0.13060 0.16810 0.32150
Group sample size:  3
Number of groups:  14
Center of group statistics:  0.1305676
Standard deviation:  0.1533338
Control limits:
 LCL      UCL
   0 0.343664
```



R 管理図

郡内範囲 R_i に関する管理図で、各群内の範囲 $R_i, i = 1, 2, \cdots, k$ を打点し、中心線 (CL)、上部管理限界線 (UCL)、下部管理限界線 (LCL) をそれぞれ

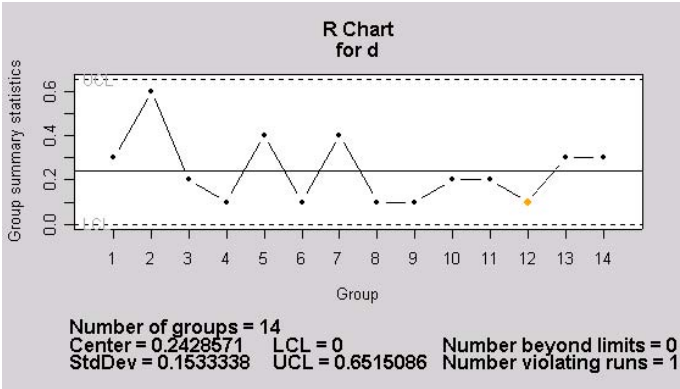
$$\mathrm{CL} = \frac{\sum_{i=1}^{k} R_i}{k} = \overline{R}, \quad \mathrm{UCL} = D_4\overline{R}, \quad \mathrm{LCL} = D_3\overline{R}$$

によって与える。ただし、群内の標本数 n が 6 以下のとき、下部管理限界線 (LCL) は考えない。ここで、係数 D_3、D_4 は右の表で与えらる。

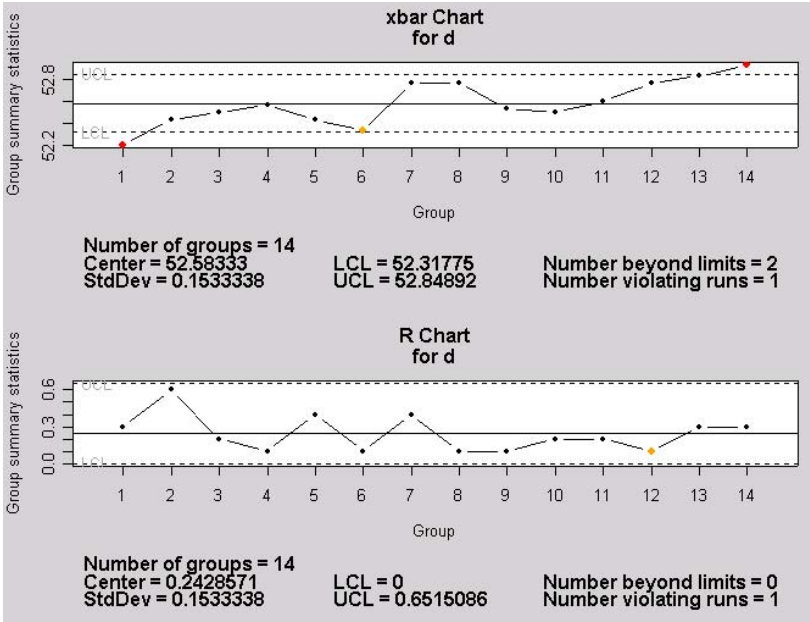
群の大きさ (k)	D_3	D_4
2	-	3.267
3	-	2.575
4	-	2.282
5	-	2.115
6	-	2.004
7	0.076	1.924
8	0.136	1.864
9	0.184	1.816
10	0.223	1.777

入力と出力結果

```
> qcc(d, type="R")
Call:
qcc(data = d, type = "R")
R chart for d
Summary of group statistics:
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.1000  0.1000  0.2000  0.2429  0.3000  0.55000
Group sample size:  3
Number of groups:  14
Center of group statistics:  0.2428571
Standard deviation:  0.1533338
Control limits:
 LCL        UCL
   0 0.55515086
```



通常は $\bar{x}$ 管理図と R 管理図を並べて描画したものがよく用いられるが、ライブラリ qcc には用意されていないので、この二つの管理図を並べることになる。



入力：出力結果は左図

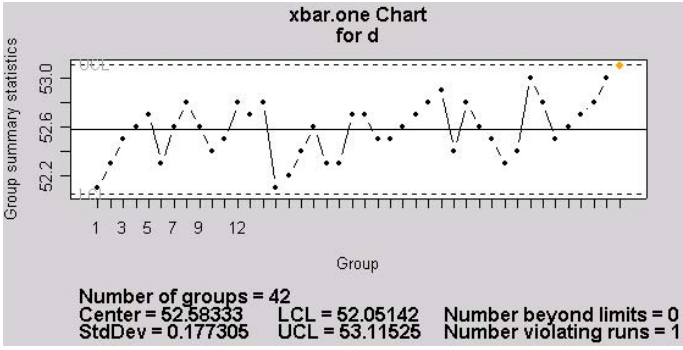
```
par(mfrow=c(2,1))
qcc(d, type="xbar")
qcc(d, type="R")
par(mfrow=c(1,1))
```

x 管理図

連続変量に対する群内の変動をみるために用いる管理図で、下記の例は体重データを一つ群として考えて扱ったものである。

入力と出力結果

```
> qcc(data = d, type = "xbar.one")
Call:
qcc(data = d, type = "xbar.one")
xbar.one chart for d
Summary of group statistics:
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  52.10   52.40   52.60   52.58   52.77   53.10
Group sample size:  3
Number of groups:  42
Center of group statistics:  52.58333
Standard deviation:  0.1773050
Control limits:
      LCL      UCL
 52.05142 53.11525
```



8.2.2　離散型の場合

生産される製品の不良個数・不良率、板の上に塗装を行ったときに発生する欠点数・欠点率等に関する計数型の管理図を考える。

p 管理図

不良率を調べるために大きさ n_i のロット (群) を考え、そのロット中の不良品の数を x_i で表す。このとき群ごとの不良率 $x_i/n_i, i = 1, 2, \cdots, k$ を打点し、中心線 (CL)、上部管理限界線 (UCL)、下部管理限界線 (LCL) をそれぞれ

$$\mathrm{CL} = \frac{\sum_{i=1}^{k} x_i}{\sum_{i=1}^{k} n_i} = \overline{p} = \text{全体での不良率}, \quad \mathrm{UCL} = \overline{p} + 3\sqrt{\frac{\overline{p}(1-\overline{p})}{n_i}}, \quad \mathrm{LCL} = \overline{p} - 3\sqrt{\frac{\overline{p}(1-\overline{p})}{n_i}}$$

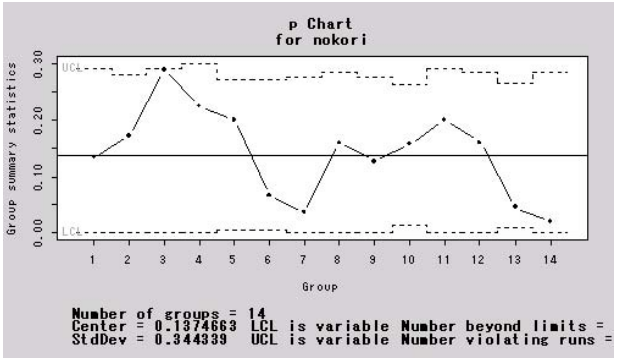
によって与える。ここで、UCL と LCL は群ごとに異なり、$\overline{p} - 3\sqrt{\frac{\overline{p}(1-\overline{p})}{n_i}} < 0$ ならば LCL は描かない。

例題 8.3. 二週間にわたり、仕入れ数と売れ残り数を調べたところ下記のようになった。売れ残った数を不良品と考えたときの p 管理図を作成せよ。■

日	1	2	3	4	5	6	7	8	9	10	11	12	13	14
仕入れ数	45	52	45	40	60	60	55	50	55	70	45	50	65	50
売れ残り数	6	9	13	9	12	4	2	8	7	11	9	8	3	1

入力と出力結果

```
> siire<- c(45,52,45,40,60,60,55,50,55,70,45,
 50,65,50)
> nokori<- c(6,9,13,9,12,4,2,8,7,11,9,8,3,1)
> qcc(nokori, sizes=siire, type="p")
Call:
qcc(data = nokori, type = "p", sizes = siire)
p chart for nokori
Summary of group statistics:
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.02000 0.08182 0.15860 0.14240 0.19330 0.28890
Summary of group sample sizes:
  sizes  40 45 50 52 55 60 65 70
  counts  1  3  3  1  2  2  1  1
Number of groups:  14
Center of group statistics:  0.1374663
Standard deviation:  0.3443390
Control limits:
         LCL       UCL
 0.000000000 0.2914594
 0.000000000 0.2807200
                      ⋮
 0.000000000 0.2914594
 0.000000000 0.2835570
 0.009336332 0.2655963
 0.000000000 0.2835570
```



pn 管理図

上の p 管理図とは異なり、各群の大きさ n は一定で、群ごとの不良個数 $x_i, i = 1, 2, \cdots, k$ を打点し、中心線 (CL)、上部管理限界線 (UCL)、下部管理限界線 (LCL) をそれぞれ

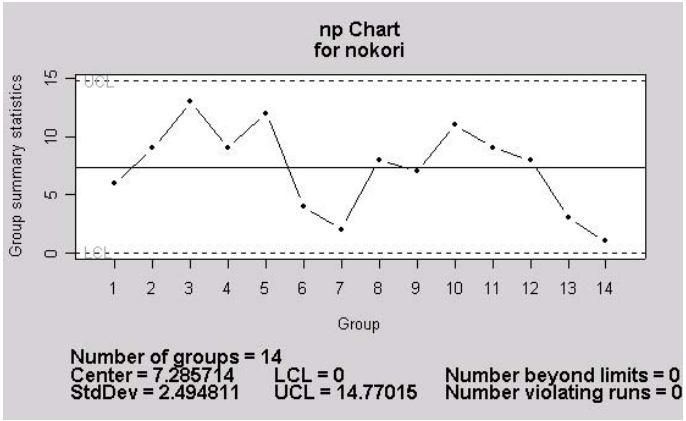
$$\mathrm{CL} = \frac{\sum_{i=1}^{k} x_i}{kn} n = \overline{p}n = \frac{\text{全不良個数}}{\text{群の数}}, \quad \mathrm{UCL} = \overline{p}n + 3\sqrt{\overline{p}n(1-\overline{p})}, \quad \mathrm{LCL} = \overline{p}n - 3\sqrt{\overline{p}n(1-\overline{p})}$$

によって与える。ここで、UCL と LCL は群ごとに異なり、$\overline{p}n - 3\sqrt{\overline{p}n(1-\overline{p})} < 0$ ならば LCL は描かない。

例題 8.4. 例題 8.3 の売れ残り数を用いる。ただし、仕入れ数は毎日一定の 50 個であったとする。このとき、*pn* 管理図を作成せよ。■

入力と出力結果

```
> qcc(nokori, sizes=siire, type="np")
Call:
qcc(data = nokori, type = "np", sizes = siire)
np chart for nokori
Summary of group statistics:
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.000   4.500   8.000   7.286   9.000  13.000
Group sample size:  50
Number of groups:  14
Center of group statistics:  7.285714
Standard deviation:  2.494811
Control limits:
 LCL      UCL
   0 14.77015
```



c 管理図

あらかじめ決まられた大きさ n の群に現れる欠点数 $c_i, i = 1, 2, \cdots, k$ を打点し、中心線 (CL)、上部管理限界線 (UCL)、下部管理限界線 (LCL) をそれぞれ

$$\mathrm{CL} = \frac{\sum_{i=1}^{k} c_i}{k} = \bar{c} = \frac{\text{全欠点数}}{\text{群の数}}, \quad \mathrm{UCL} = \bar{c} + 3\sqrt{\bar{c}}, \quad \mathrm{LCL} = \bar{c} - 3\sqrt{\bar{c}}$$

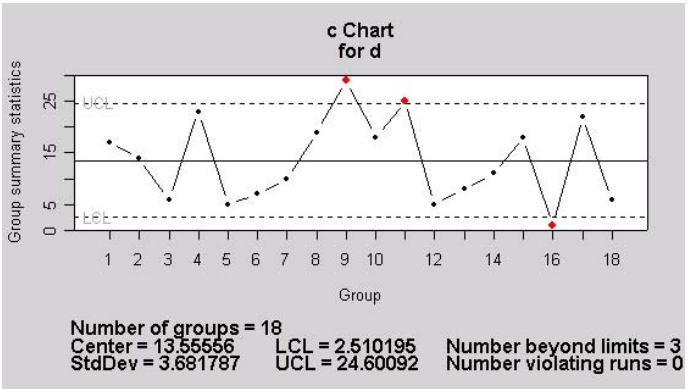
によって与える。ここで、$\bar{c} - 3\sqrt{\bar{c}} < 0$ ならば LCL は描かない。

例題 8.5. 1 日に 30 個の製品を製造しており、18 日間の塗装面の欠点数を数えたところ下記の表を得た。*c* 管理図を作成せよ。■

日	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
欠点数	17	14	6	23	5	7	10	19	29	18	25	5	8	11	18	1	22	6

入力と出力結果

```
> qcc(d, type="c")
Call:
qcc(data = d, type = "c")
c chart for d
Summary of group statistics:
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1.00    6.25   12.50   13.56   18.75   29.00
Group sample size:  1
Number of groups:  18
Center of group statistics:  13.55556
Standard deviation:  3.681787
Control limits:
      LCL      UCL
 2.510195 24.60092
```



u 管理図

c 管理図を一般化したもので、各群 (単位) の大きさは n_i と群ごとに異なっていてもよく、単位あたりの欠点数 $\frac{c_i}{n_i}, i = 1, 2, \cdots, k$ を打点し、中心線 (CL)、上部管理限界線 (UCL)、下部管理限界線 (LCL) をそれぞれ

$$\mathrm{CL} = \frac{\sum_{i=1}^{k} c_i}{\sum_{i=1}^{k} n_i} = \bar{u} = \frac{\text{全欠点数}}{\text{全単位数}}, \quad \mathrm{UCL} = \bar{u} + 3\sqrt{\frac{\bar{u}}{n_i}}, \quad \mathrm{LCL} = \bar{u} - 3\sqrt{\frac{\bar{u}}{n_i}}$$

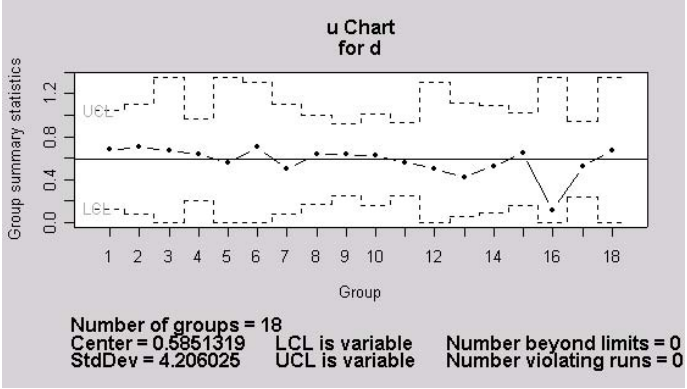
によって与える。ここで、UCL と LCL は群ごとに異なり、$\bar{u} - 3\sqrt{\frac{\bar{u}}{n_i}} < 0$ ならば LCL は描かない。

例題 8.6. 例題 8.5 の欠点数のデータを用いる。ただし、製品の個数は日ごとに変動するものとする。u 管理図を作成せよ。■

日	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
製品数	25	20	9	36	9	10	20	30	46	29	45	10	19	21	28	9	42	9
欠点数	17	14	6	23	5	7	10	19	29	18	25	5	8	11	18	1	22	6

入力と出力結果

```
> qcc(d, sizes=size, type="u")
Call:
qcc(data = d, type = "u", sizes = size)
u chart for d
Summary of group statistics:
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.1111  0.5238  0.55256  0.5706  0.55607  0.7000
Summary of group sample sizes:
  sizes  9 10 19 20 21 25 28 29 30 36 42 45 46
  counts 4  2  1  2  1  1  1  1  1  1  1  1  1
Number of groups:  18
Center of group statistics:  0.5851319
Standard deviation:  4.206025
Control limits:
        LCL       UCL
 0.12616841 1.0440954
 0.07199512 1.0982687
 0.00000000 1.3500710
                      ⋮
 0.23103388 0.9392299
 0.00000000 1.3500710
```



8.3　チェックシート

対象についての調査項目、記録すべき事柄を表の形に配置したもので、これにより必要なデータが容易に収集・記録が出来るようになるものである。また、作業の点検・確認を行うことを目的とするチェックシートもあり、実際の図・写真などを入れることで点検作業を円滑に行うことが出来るようになる。

8.4　層別

層別とは、データを工場別、担当者別、機械・設備別、原材料別、作業方法別などに分類することである。これは製品の仕上がりにはバラツキがあり、このバラツキの原因を見つけ出そうとする際、層別することにより、工場間・担当者間で異なる歩留まり率などを引き起こす要因を見つけることが容易になり、品質を保った生産に役立つ。

新 QC 七つ道具

1. 親和図法
2. 連関図法
3. 系統図法
4. マトリックス図法
5. アロー・ダイアグラム法
6. PDPC 法
7. マトリックス・データ解析法

と呼ばれているものも提案されている。いずれのの手法を用いたにせよ、生産工程の改善を行ったことによりどのような効果が実際にあがったかを確認をすることは、品質の維持・管理を進めていくには重要なことです。

付録 A　検索

A.1　インターネット検索

インターネットから情報を得る方法として、系統的にまとめられているものとして、「オンライン」百科事典である “ウィキペディア” を利用する方法と検索サイトから適当なホームページを探し出す方法がある。

A.1.1　Wikipedia の利用

ウィキペディア (Wikipedia)(http://ja.wikipedia.org/) は利用している者によって、項目が加筆・修正されていく百科事典である。

検索の欄に検索項目を入力し、検索ボタンをクリックするか Enter キーを押すことで系統だった情報を手に入れることが出来ます。この Web Page はお気に入り (Bookmarks) に入れておくと重宝します。利用者の加筆・修正ということで日本語の Page が充実していない箇所も見られますが、英語や他言語で記述された Page へのリンクが張られており、他言語での表現も調べられます。

A.1.2　検索サイトの利用

goo(http://www.goo.ne.jp) を利用したインターネット検索について考える。

他の検索サイトとして、google(http://www.google.co.jp)、Yahoo!(http://www.yahoo.co.jp) 等があるが、英和、和英、国語辞典等の辞書機能が備わっているという点で、ここでは検索サイト goo の利用について述べる。辞書に関しては、ネット接続が出来ない場合に備え、英辞郎[*1]や広辞苑等の電子辞書が利用できる環境を作っておくことも必要かもしれません。

インターネット検索を利用する際、検索欄に用語を一つだけ入力しても欲しい情報が得られる場合はそれほど多くない。そこで、以下の機能を利用して検索すると目的の情報が比較的簡単に得られる。

- 用語の定義を調べる。
  用語欄に調べたい用語の後に “とは” を付けて検索する。
  例　「統計学とは」。
- AND 検索
  用語欄に「用語 1␣AND ␣用語 2…」の形で “␣AND ␣” を用語間に入れて検索する。
  例　「彦根␣食べ放題」。「彦根」と「食べ放題」という用語を含むウェブページが抽出される。
- OR 検索
  用語が 2 種類以上の形で表現されるとき等に用いる。
  例　「彦根␣(食べ放題␣OR␣食い放題)」。上の AND 検索に比べ、「食い放題」という言葉を含んだウェブページが抽出される。
- NOT 検索
  – を用いて、除外したい用語のあるウェブページを抽出しない。
  例　「いとしのエリー – サザン’」。「いとしのエリー」を含むが、「サザン」を含まないウェブページが抽出される。
- フレーズ検索 (""を用いて検索する)。
  例　「"black and white"」と入力し、black and white というフレーズがこの順序で含まれるウェブページを抽出する。「"Solaris 10"」。
- サイトを指定して検索する。
  例　「奨学金 site:www.biwako.shiga-u.ac.jp」。滋賀大学経済学部 www.biwako.shiga-u.ac.jp のサイトから “奨学金” を含むページを抽出する。
- 辞書検索
  調べたい言葉を入力し、英和、和英、国語などの辞書のボタンをクリックします。英語辞書を用いた検索では代表的な単語についての音声ファイルがあります。
- ファイルの拡張子を指定。
  例　「発表␣filetype:ppt」。“発表” を含み、filetype でアプリケーションの形式 (拡張子) を指定し、文書をを抽出する。PowerPoint の拡張子[*2]が ppt であるからこれを付けて検索。
  代表的な拡張子は以下のものである。

アプリケーション	Word	Excel	PowerPoint	Acrobat (Reader)
拡張子	doc	xls	ppt	pdf

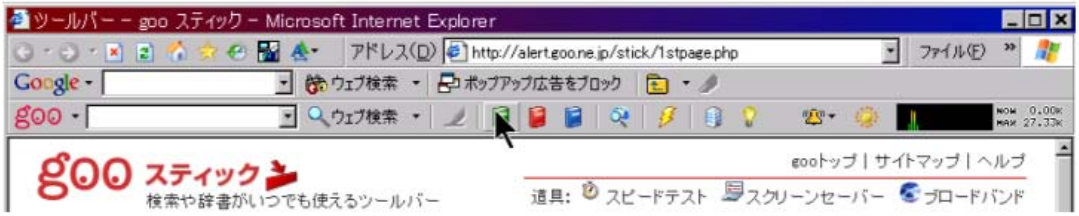
AND 検索、OR 検索、NOT 検索は goo の検索画面にある “検索オプション”

[*1] 英辞郎については http://www.eijiro.jp/ 参照。Mac OS X では英辞郎ビューアー (http://numata.aquasky.jp/software/eview/)、Windows では Personal Dictionary for Win32(http://homepage3.nifty.com/TaN/) を利用します。英辞郎と Personal Dictionary for Win32 は Shareware で、英辞郎ビューアは Freeware です。

[*2] Web サイト：拡張子辞典 http://www.jisyo.com/viewer/ 参照。

を利用して検索することが出来ます。

goo を用いて検索は、goo スティックをインストールし、goo のウェブページを開くことなく、直接インターネットエクスプローラーから各種検索が出来るようになります。他の検索サイトも同様の検索バーがあります。



上のものは google バーも入れてみました。goo スティックのオプションで表示項目、検索結果の表示方法、検索ページの言語指定等が出来ます。

R-Tips　閲覧ソフトの操作について

- 検索結果が表示されたとき、Shift キーを押しながら、結果の url をクリックすることで別ウィンドウに表示してくれます。
- 結果のホームページが表示されたら、Windows のときは Ctrl キー (Mac のときはコマンドキー) を押しながら F を押すことでページ内の検索機能が利用できます。
- 結果のホームページが表示されないときは goo のキャッシュ機能を用います。

[goo を用いて "FireFox" を検索したときの一番目に現れた項目]

これは goo の検索ソフトがデータを収集したときの結果を goo が記憶していたものです。

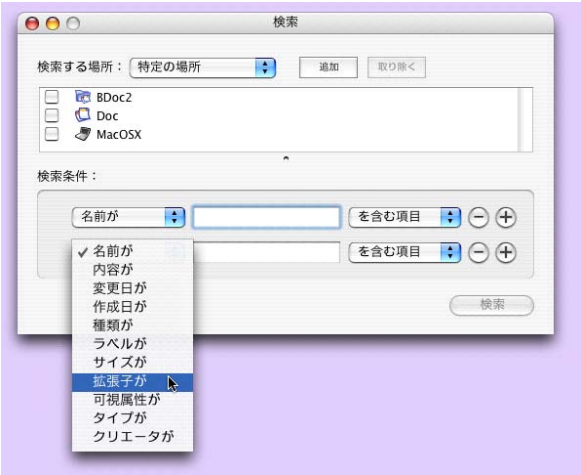
- 関心のあるページが現れたら、"お気に入り"(Bookmarks) に追加しておきましょう。これはメニューのお気に入りから追加を選択することで、そのページの url が保存されます。
- お気に入りを他の機械に移すことは可能で、メニューの "ファイル" の "インポートおよびエクスポート" を選択し、お気に入りをファイルとして書き出すことができます。この書き出されたファイルを USB メモリなどにコピーし、他の機械で読み取ります。また、メールの添付ファイルとして他の機械・他の方に受け渡すことも出来ます。この書き出されたファイルは html 形式になっており、クリックすることで一覧が表示されます。インターネットエクスプローラーではお気に入りの順序が保たれた形でエクスポートされないので、FireFox などの他の閲覧ソフト[*3]を利用するとお気に入りが順序を保ったままの形で他のソフトに移行できます。

A.2　HDD 内の検索

利用しているパソコン内のデータ検索を考えます。"PC 内␣検索ソフト" という用語でインターネット検索を行うと、いろいろなソフトがあることが分かりますが、ここでは利用している基本ソフト (OS:Operating System) に組み込まれている機能の説明を与えます。

A.2.1　Mac OS X の場合

Finder に切り替え、コマンドキーを押しながら、F のキーを押します。ファイルのある HDD の前にチェックマークを付けます。



- ファイル名が分かっているとき
  ファイル名を入力し、Enter キー↵を押します。
- ファイル名が分からないが拡張子は分かっているとき
  “名前” の代わりに拡張子を選択し、拡張子を入力し、return キーを押します。
- その他
  - ファイル名入力欄の隣の + をクリックし、検査項目を増やし、ファイルを絞り込みます。
  - 特定のフォルダーにあることが分かっているときは、フォルダーを検索する場所にドラッグ& ドロップし、検索範囲を縮小させることが出来ます。

A.2.2　Windows の場合

文章、画像が入っていると思われる Folder を開けます。文章は、作成したソフト (アプリケーション) により、Word なら拡張子が doc、Excel なら xls と自動的に付加され、Windows のシステムによりソフトとの関連付けがなされています。画像の拡張子は gif、jpg、bmp、png 等が付けられており、閲覧ソフトで表示することが出来ます。

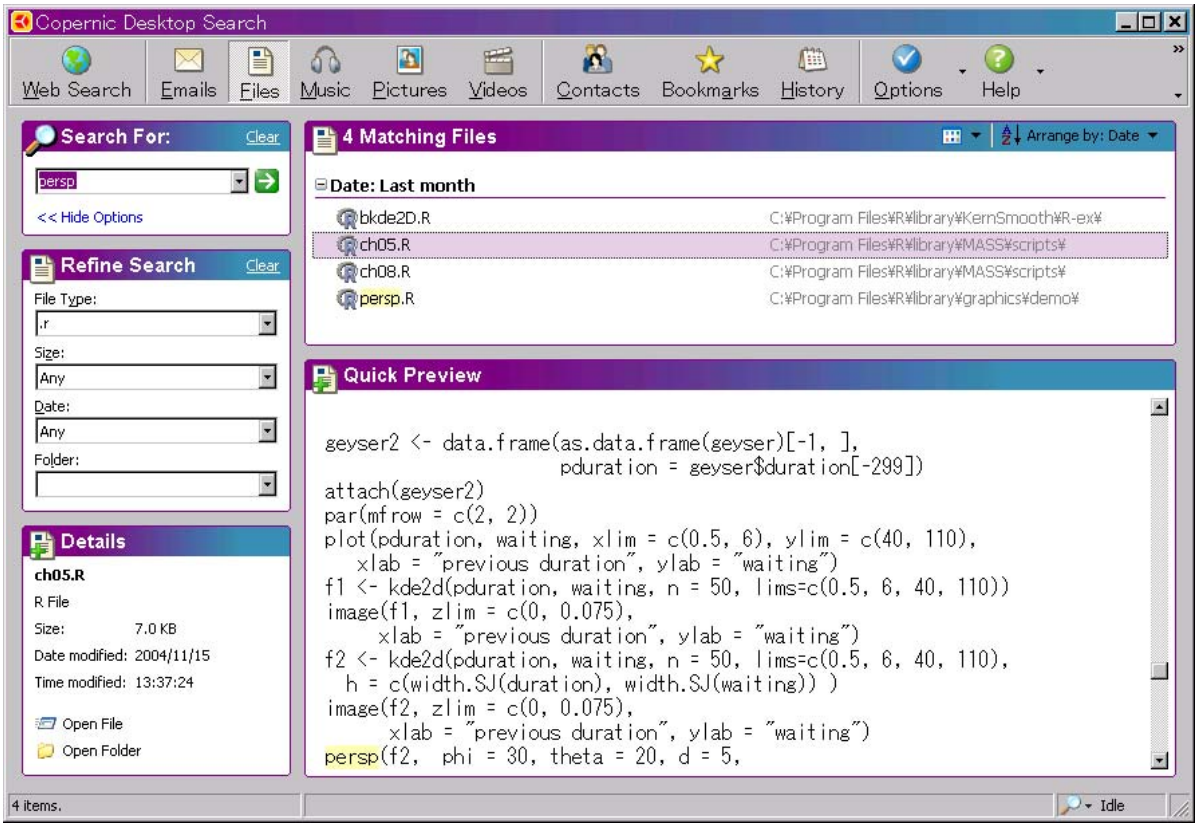
Folder が開いている状態で、Ctrl キーを押しながら、F のキーを押します。

- ファイル名が分かっているとき
  “ファイル名のすべてまたは一部” の所に、ファイル名を入れ、検索ボタンをクリックします。Enter キーを押しても検索が始まります。
- ファイル名が分からないが拡張子は分かっているとき
  “ファイル名のすべてまたは一部” の所に、半角の「*.」と拡張子を入れ (Word 文書を探すときは、半角で「*.doc」を入力)、検索ボタンをクリックします。
- 拡張子が分かっており、文章の内容も少し分かっているとき
  “ファイル名のすべてまたは一部” の所に、半角の「*.」と拡張子を入れ、その下の “ファイルに含まれる単語または句” にファイルに含まれている言葉を入れ、検索します。

- その他
  変更時期等が分かっているときは、“いつ変更されましたか？” に変更時期を入力し、検索します。

Tips　HDD 内の検索ソフトについて

Windows において HDD 内のファイル検索には Copernic Desktop Search(http://www.copernic.com/en/products/desktop-search/) などのソフトを用いると検索したいフォルダーを複数指定でき、Windows に組み込まれている検索より、速く検索一覧を得ることが出来ます。また、検索一覧の下部にプレビュー画面もあり、ファイルをすべて開くことなく目的のファイルに到達できます。pdf ファイルの内容も検索してくれます。下の例は R をインストールしたフォルダ内の文書で、persp を含む拡張子.r の文書を探しています。



上の図で、拡張子.r のアイコンは R のアイコンに変更してあります。

付録 B　仮説検定について

硬貨を 10 回投げ、表が 1 回しか出なかったという状況を考える。この硬貨は公正な歪みのない硬貨 (fair coin) であろうか？硬貨を 10 回投げたとき表の出る回数を確率変数 X で表し、表の出る確率を p としたとき、X はパラメータ (n, p) の二項分布に従い、10 回投げて k 回表が出る確率は

$$p_k = \Pr\{X = k\} = {}_{10}\mathrm{C}_k\, p^k q^{10-k}, \qquad k = 0, 1, 2 \cdots, 10$$

によって与えられる。ここで、q は裏の出る確率 $1 - p = q$ である。

このとき、表の出る確率が $\frac{1}{2}$ という仮説の下で、表の出る回数 k についての確率は右表のように計算され[*1]、表の出る回数が 1 回以下となる確率は 0.01074219 である。このとき、「表が 1 回しか出ないということでこの硬貨は歪んでいる」と判断しても差し支えないだろうか？歪んでいるか歪んでいないかは表の出る確率 p が $\frac{1}{2}$ であるかいなかで測るものとする。

k	確率 (p_k)	累積
0	0.000976563	0.000976563
1	0.009765625	0.01074219
2	0.04394531	0.0546875
3	0.1171875	0.171875
4	0.2050781	0.3769531
5	0.2460938	0.55230469
6	0.2050781	0.828125
7	0.1171875	0.9453125
8	0.04394531	0.9892578
9	0.009765625	0.9990234
10	0.000976563	1

歪んでいない ($p = \frac{1}{2}$) という仮定の下で、表が 1 回以下の確率は 0.000976563 + 0.009765625 = 0.01074219 であり、これは歪んでいない硬貨を 10 回投げることを 100 回繰り返したら、1 回位起こることを意味している。硬貨は歪んでいるかということを考えているので、表の出る回数が多くても歪んでいるという判断を下す必要がある。よって、表が 9 回以上の出たときも歪んでいると考えると、この確率は 0.01074219 となり、表の出た回数が 1 回以下の確率と 9 回以上が起こる確率をあわせると、0.0107 + 0.0107 = 0.0214 となる。この確率の低さから、硬貨の表の出た回数が 1 回ということは滅多に起こらないことが起きているのでなく、硬貨は歪んでいると考えるのが自然なことに思われる。しかし、公正な硬貨を投げたとしても 1 回以下または 9 回以上表が出る確率は 0.0214 でゼロではないから、公正な硬貨を投げているにもかかわらず硬貨を歪んでいると判断する過ちが発生する。

「歪んでいない」という仮説を**帰無仮説**(null hypothesis) といい、「硬貨が歪んでる」という仮説を**対立仮説**(alternative) という。また、表が 1 回以下または 9 回以上出る確率 0.0214 という値は帰無仮説が正しいにも関わらず帰無仮説を受け入れない (棄却する) 確率で、これを**有意水準**(significance level) と呼ぶ。この α の値はあまり大きくても帰無仮説を間違って棄却するということを導くので、安全性、信頼性により変化するが、通常、0.05 あるいは 0.01 にとることが多い。

上の検定問題に対しては

表の出た回数が 1 回以下 ($X \leq 1$) または 9 回以上 ($X \geq 9$) ならば、帰無仮説を棄却する。
表の出た回数が 2 回以上 8 回以下 ($2 \leq X \leq 8$) ならば、帰無仮説は棄却しない。

という検定方式を考えた。このとき、$X \leq 1$ または $X \geq 9$ という X の結果の範囲を検定問題の**棄却領域**(rejection region) という。歪んでいない硬貨を投げるとき、$X \leq 1$ となる確率は 0.01074219 で $X \geq 9$ となる確率も 0.01074219 で合計 0.0214 であるので、有意水準は 0.0214 となる。

他の例として、不良率 0.3% の工場の製品の 100 個について調査したところ、4 個の不良品があったと仮定する。不良率 0.3% というのは過去の経験から信じられている値で、4 個の不良品があったということは不良率が 0.3% より大きいのではないかと思われる。このとき、帰無仮説は

$$\mathcal{H}_0 : \text{本当の不良率 } p \text{ は } 0.003 \text{ に等しい。}$$

となり、対立仮説は

$$\mathcal{H}_1 : \text{本当の不良率 } p \text{ は } 0.003 \text{ より大きい。}$$

となる。このとき、帰無仮説の下で不良品の個数を確率変数 X で表すと、

$$\Pr\{X = k\} = {}_{100}\mathrm{C}_k 0.003^k \times 0.997^{100-k}, \quad k = 0, 1, 2 \cdots, 100$$

[*1] `p<- 1/2;p10<- p^10;k<- 0:10;choose(10,k)*p10` により、表の値は得られている。

となり、不良品が多いとき帰無仮説を棄却することになる。

k	確率 (q_k)
0	1
1	0.2595157
2	0.03670202
3	0.003514622
4	0.0002524704
5	0.0000144347
6	0.0000068262

実際、どの程度不良品の個数があったとき帰無仮説を棄却すればいいのかが問題となる。左の表は k の値に対して

$$
\begin{aligned}
q_k = \Pr\{X \geq k\} &= \sum_{i=k}^{100} \Pr\{X = i\} \\
&= \sum_{i=k}^{100} {}_{100}\mathrm{C}_i \, 0.003^i \times 0.997^{100-i} \qquad \text{(B.1)}
\end{aligned}
$$

をまとめたもので、帰無仮説の値 $p = 0.003$ が正しいとしたとき不良品の個数が k 個以上となる確率を表している。

上の表の確率 q_k は、右に与えられる R の関数を用いて計算した。1 行目はオブジェクト p に確率 p の値 0.003 を代入し、確率 $1-p$ を q とし、k を 0 から 100 までの数列とする。2 行目は式 (B.1) の総和の中の確率を計算している。3 行目ではそれぞれ $q_0, q_1, \cdots, q_6$ の確率を計算している。4 行目から 10 行目はその結果である。

入力

```
1 :   > p<- 0.003; q<- 1-p; k<- 0:100
2 :   > prob<- choose(100,k)*p^k*q^(100-k)
3 :   > for (i in 1:6) print(sum(prob[i:100]))
4 :   [1] 1
5 :   [1] 0.2595157
6 :   [1] 0.03670202
7 :   [1] 0.003514622
8 :   [1] 0.0002524704
9 :   [1] 1.443473e-05
10 :  [1] 6.826204e-07
```

不良率 0.3% の工場において、左上の確率の表から不良品が 1 個以上発生する確率は 0.259 で、2 個以上の確率は 0.037 である。3 個以上不良品が発生する確率は 0.0035 なので今の場合 4 個発生しているので、有意水準を 5% としたとき帰無仮説は棄却されることになる[*2]。

いままで述べてきたことから、標本に基づき判定を行う際、真の状態と判定との関係が生じる。

母集団の状態 ＼ 判定	$\mathcal{H}_0$ を棄却しない	$\mathcal{H}_0$ を棄却
帰無仮説 $\mathcal{H}_0$ が真	正しい判定	誤った判定（第一種の過誤）
対立仮説 $\mathcal{H}_1$ は真	誤った判定（第二種の過誤）	正しい判定

左の表において

(1) 帰無仮説 $\mathcal{H}_0$ が真のとき、帰無仮説 $\mathcal{H}_0$ を棄却しない。
(2) 対立仮説 $\mathcal{H}_1$ が真のとき、帰無仮説 $\mathcal{H}_0$ を棄却する。

は正しい判定であるが、

(3) 帰無仮説 $\mathcal{H}_0$ が真のとき、帰無仮説 $\mathcal{H}_0$ を棄却する。
(4) 対立仮説 $\mathcal{H}_1$ が真のとき、帰無仮説 $\mathcal{H}_0$ を棄却しない。

は誤った判定である。(3) の場合の**第一種の過誤**(Type I Error) は、帰無仮説が正しいにもかかわらず帰無仮説を棄却してしまう誤りで、“ 生産者危険 ”、“ あわて者の誤り ” とも呼ばれ、この誤りをおかす確率を有意水準 (significance level) と呼び、通常ギリシャ文字の α で表すことが多い。(4) の場合の**第二種の過誤**(Type II Error) は、帰無仮説が真でない状況において帰無仮説を棄却しない誤りで、“ 消費者危険 ”、“ ぼんやり者の誤り ” とも呼ばれ、この誤りをおかす確率は通常ギリシャ文字の β で表される。また、$1-\beta$ は検定の**検出力**(power) と呼ばれ、帰無仮説が真でないときに真でないと判定できる確率を表している。

上で現れた用語の再確認として、下記の検定問題を考える。

大きさ n の標本 $X_1, X_2, \cdots, X_n$ が正規分布 $\mathcal{N}(\mu, 1^2)$ から抽出され、有意水準 5% で

$$\text{帰無仮説 } \mathcal{H}_0 : \mu = 0 \ \textit{versus} \ \text{対立仮説 } \mathcal{H}_1 : \mu > 0$$

を検定する。

[*2] 有意水準が 5% なので不良品が 2 個でも帰無仮説は棄却されている

この検定問題に対して、標本から計算される標本平均

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

を用いて検定を行うことになる。$\overline{X}$ を利用することは尤度比検定から導かれることで、この検定問題に対して他の検定統計量より優れている (一様最強力検定) ことが示されている。これは、データ一つ一つを用いて検定を考えるより、集約した標本平均 $\overline{X}$ を用いた方がよく、他の統計量より標本平均が優れているということである。

さらに、正規分布の再生性から

$$\overline{X} \sim \mathcal{N}\left(\mu, \left(\frac{1}{n}\right)^2\right)$$

が成り立つ。この性質と対立仮説の形から $\overline{X}$ の値が大きいとき帰無仮説を棄却するという検定方式が考えられる。定数 C を用いて棄却領域を $(\overline{X} > C)$ と表すと、有意水準が α であるから定数 C は

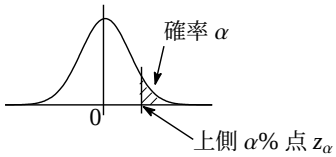
$$\Pr\{\overline{X} > C\} = \alpha$$

を満たす必要がある。従って、Z を標準正規分布に従う確率変数とすると、

$$\begin{aligned}\alpha = \Pr\{\overline{X} > C\} &= \Pr\left\{\frac{\overline{X} - 0}{\sqrt{1/n}} > \frac{C - 0}{\sqrt{1/n}}\right\} \\ &= \Pr\left\{Z > C\sqrt{n}\right\} \\ &\therefore \quad C = z_\alpha / \sqrt{n}\end{aligned}$$

となる。ここで、z_α は標準正規分布の上側 α% 点で、代表的な α の値に対して統計書、教科書の巻末の表として掲載されている[*3]。ソフトウェア R では `qnorm(1-α)` として計算される。つまり、z_α は標準正規分布 $\mathcal{N}(0, 1^2)$ に対して

$$\alpha = \int_{z_\alpha}^{\infty} \frac{1}{\sqrt{2\pi}} \exp(-\frac{x^2}{2})\, dx$$

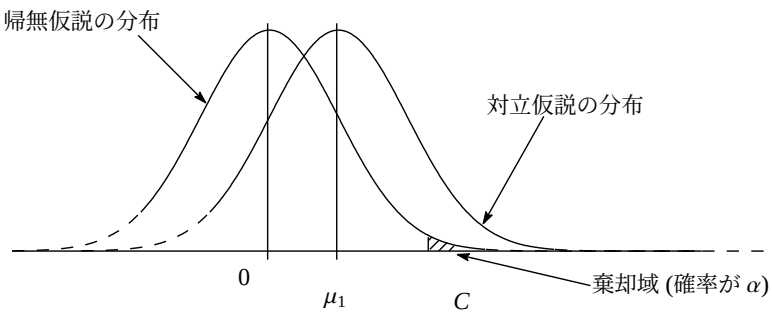


である。

よって、いまの検定問題に対する棄却領域は

$$\overline{X} > \frac{z_\alpha}{\sqrt{n}}$$

となる。この領域に入る標本平均 $\overline{X}$ に対して帰無仮説 $\mathcal{H}_0$ を棄却することになる。

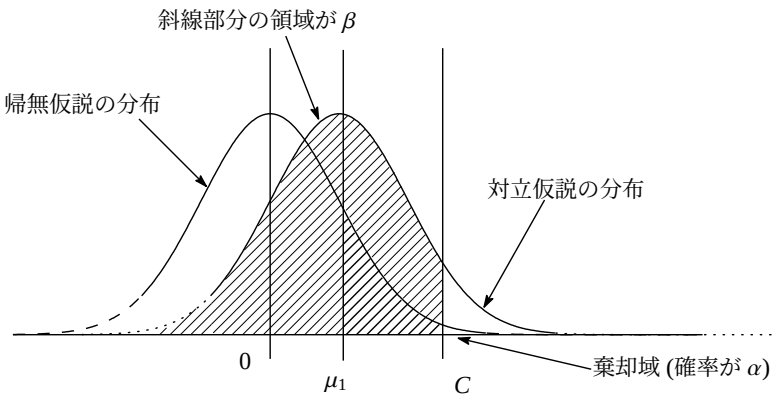


検出力 $1-\beta$ を考えるために対立仮説の一つの値を μ_1 とする。ここで、対立仮説は $\mu > 0$ という形であるが、一つの値を指定しないと検出力は計算できないので、ここではその値を μ_1 とおく。第二種の過誤の確率は対立仮説の一つの値 μ_1 が真であるとき帰無仮説を棄却しないものであったので、この確率は平均が μ_1 のとき

$$\beta = \Pr\{\overline{X} < z_\alpha / \sqrt{n}\}$$

を計算するものである。

[*3] 教科書によっては z_α を他の意味に用いていることもあるので注意が必要です。



$\mu = \mu_1$ という対立仮説の下で

$$\overline{X} \sim \mathcal{N}\left(\mu_1, \left(\frac{1}{n}\right)^2\right)$$

であるから、標準正規確率変数 Z を用いて

$$\begin{aligned}\beta = \Pr\{\overline{X} < z_\alpha/\sqrt{n}\} &= \Pr\left\{\frac{\overline{X}-\mu_1}{\sqrt{1/n}} < \frac{z_\alpha/\sqrt{n}-\mu_1}{\sqrt{1/n}}\right\} \\ &= \Pr\left(Z < z_\alpha - \mu_1\sqrt{n}\right) \\ \therefore \quad 1-\beta &= \Pr\left(Z \geq z_\alpha - \mu_1\sqrt{n}\right)\end{aligned} \tag{B.2}$$

となる。この値は μ_1 の実際の値により、正規分布表もしくは R の関数 qnorm を用いて計算することになる。

- 有意水準の図と検出力の図より、有意水準を小さくすると、検出力が小さくなり検定の意味がなくなる
- 標本数 n が大きくなると、標本平均 $\overline{X}$ の分散 $1/n$ がゼロに近づき、検出力が増加する
- 式 (B.2) より、対立仮説の値 μ_1 が大きくなると β の値が減少、つまり検出力 $1-\beta$ が増加する
- 式 (B.2) より、標本数 n が増加すると、β の値が減少、つまり検出力 $1-\beta$ が増加する

ということが分かる。

同じように、検定問題

> 大きさ n の標本 $X_1, X_2, \cdots, X_n$ が正規分布 $\mathcal{N}(\mu, 1^2)$ から抽出され、有意水準 5% で
>
> 帰無仮説 $\mathcal{H}_0 : \mu = 0$ *versus* 対立仮説 $\mathcal{H}_1 : \mu \neq 0$
>
> を検定する。

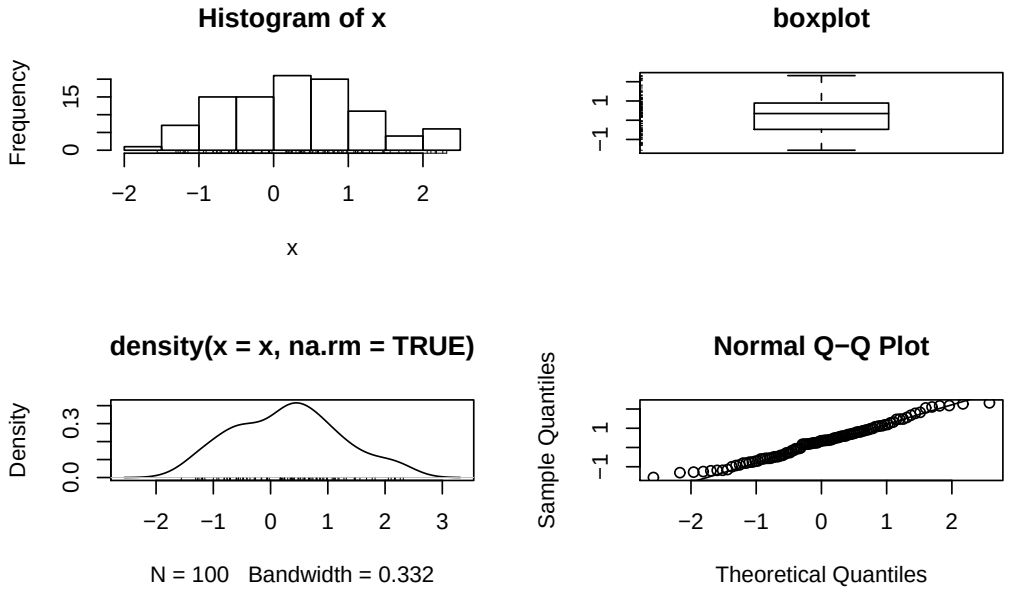
を考えると、この問題の検定方式は、有意水準 α に対して

$$|\overline{X}| > \frac{z_{\alpha/2}}{\sqrt{n}} \Longrightarrow \text{帰無仮説 } \mathcal{H}_0 \text{を棄却する。}$$
$$|\overline{X}| \leq \frac{z_{\alpha/2}}{\sqrt{n}} \Longrightarrow \text{帰無仮説 } \mathcal{H}_0 \text{は棄却できない。}$$

となる。ここで、$\Pr\{|Z| > z_{\alpha/2}\} = \alpha$ である。このときの検出力も上の場合と同じようにして求まる。

最初の問題のように対立仮説が $\mu > 0$ ときのように棄却領域が片側に設定されるとき検定を**片側検定**(one-sided test) といい、第二の問題の対立仮説が $\mu \neq 0$ ときのように棄却領域が両側に設定されるとき検定を**両側検定**(two-sided test) と呼ぶ。

母集団が正規分布と想定される検定を行う際、標本がどのようななっているかを見るために下記のようなものをプロットしておくと良い。



順に時計回りにヒストグラム[*4]、箱ひげ図[*5]、確率密度プロット、正規確率プロット[*6]である。

ヒストグラムは階級数、階級区間の取り方に一定のルールが無く、標本の性質を表しているものを必ずしも表していないということから、確率密度プロットの方が好まれる。この確率密度プロットは kernel density estimate という方法に基づいており、一般に提案されている 7 つの手法 ("gaussian"、"epanechnikov"、"rectangular"、"triangular"、"biweight"、"cosine"、"optcosine") が R には組み込まれている。上図を描くためのソフトウェア R の関数は下記のものである。

入力

```
dataA<-function (x)
{
    par(mfrow = c(2,2))
    hist(x)
    rug(x)
    boxplot(x)
    rug(x, side = 2)
    title("boxplot")
    plot(density(x, na.rm=TRUE))
    rug(x)
    qqnorm(x)
    qqline(x)
    par(mfrow=c(1,1))
}
```

[*4] 7 ページ参照。
[*5] 10 ページ参照。
[*6] 9 ページ参照。

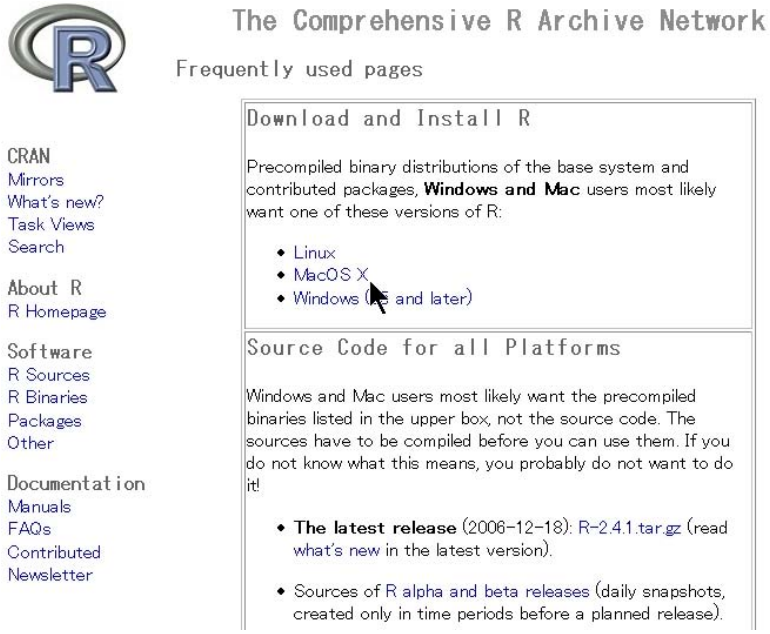
付録 C　R のインストール

The R Project for Statistical Computing(http://www.r-project.org/) から CRAN（Comprehensive R Archive Network）を選択し、日本のミラーサーバー筑波大学[*1]、東京大学[*2]または会津大学[*3]から利用しているパソコンの OS に適応したソフトウェア R をダウンロードすることから始まります。ここでは、Mac OS X 版と Windows 版のインストールの説明を行う。Unix や Linux(Debian、Mandrake、FedoraCore、SUSE、Vine) についても同様にインストールをすることができます。

C.1　Mac OS X 版

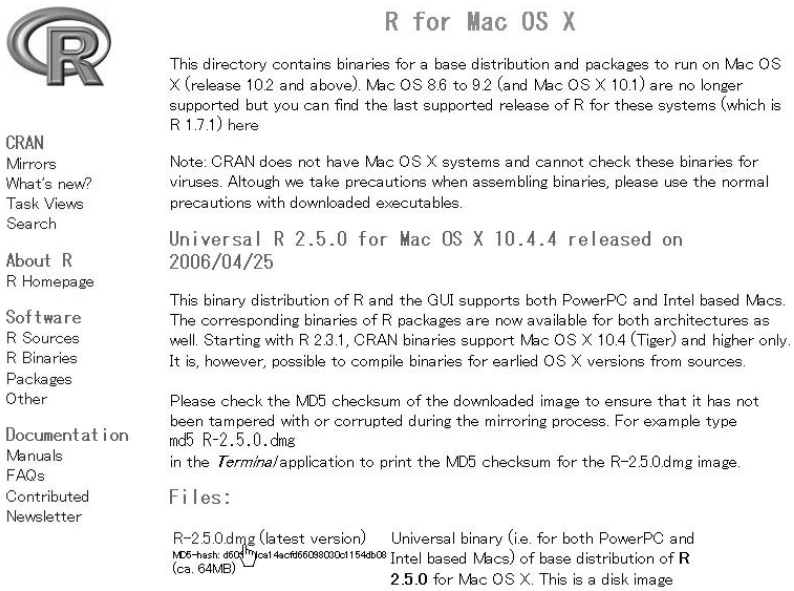
C.1.1　R のインストール

日本のミラーサーバー、例えば、筑波大学サーバーのページ



の MacOS X をクリックし、

[*1] http://cran.md.tsukuba.ac.jp/
[*2] ftp://ftp.ecc.u-tokyo.ac.jp/CRAN/
[*3] ftp://ftp.u-aizu.ac.jp/pub/lang/R/CRAN



ファイル “R-2.5.0.dmg” をクリックします。ダウンロードが始まり、解凍され、デスクトップに

が現れます。

解凍された “R.mpkg” をクリックすることで、インストールが始まり、

順に “続ける” をクリックし、使用許諾契約に同意し、

起動ディスクを選択し、“続ける” をクリックし、“インストール” をクリックします (上右画面は “インストール” でなく “アップグレード” となっています)。

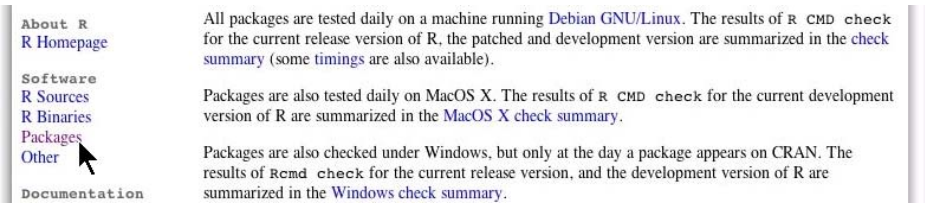
ソフトウェアが正常にインストールされると、アプリケーションのフォルダー内に

のアイコンが作成され、これをクリックすることで R が起動されます。必要なら、このアイコンをドックにドラッグ&ドロップします。

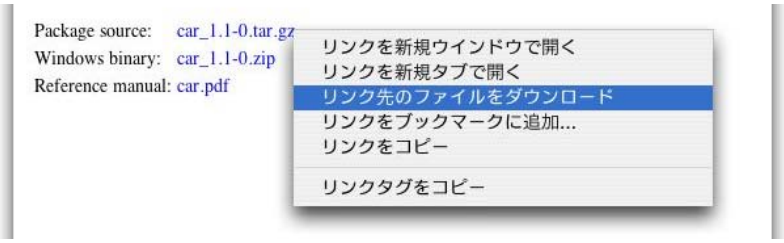
C.1.2　パッケージのインストール

HDD に tgz ファイルをダウンロードした後インストール

パッケージを HDD にダウンロードし、それを R に組み込むことを説明します。これはネットに繋がっている状態のときに必要なパッケージをダウンロードし、その後他のパソコンにもインストールするときにも利用できます。CRAN ウェブページ[*4]の左フレームの “Packages” をクリックすると、右のフレームに “Contributed Packages” が現れます。



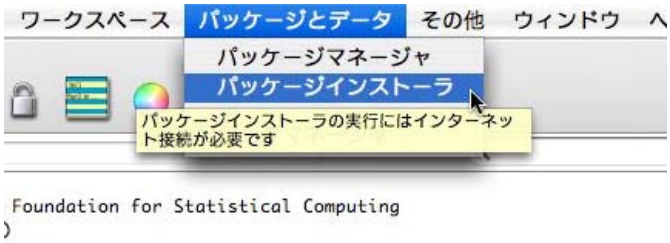
この “Contributed Packages” からダウンロードしたい Package 名 (以下の例は car の場合) を探し、Package source“car_1.1.0.tar.gz” を Ctrl キーを押しながら、クリックし “リンク先のファイルを別名でダウンロード...” を選択し、



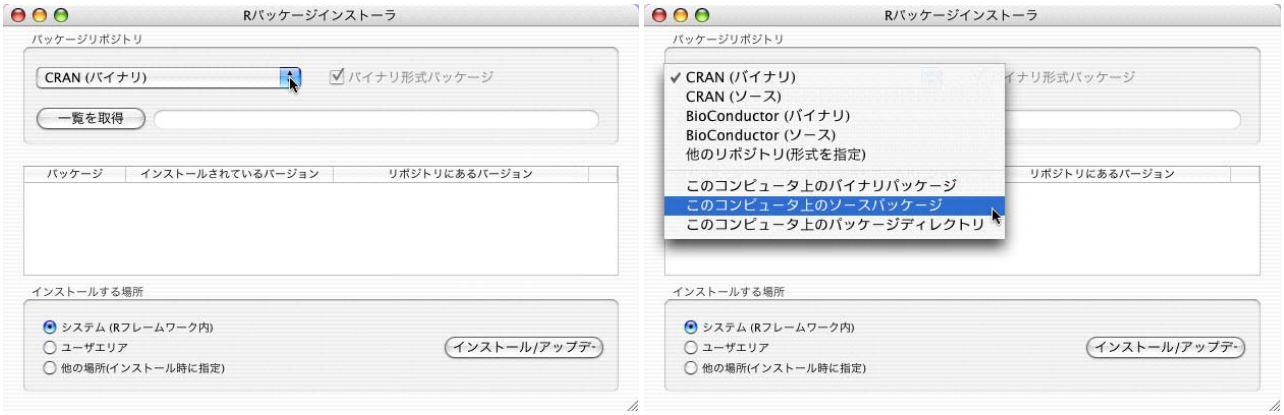
適当な所にダウンロードします。他の機械にパッケージをインストールするときは、これを USB メモリー等の外部記憶装置に保存しておきます。ファイルをダブルクリックしてダウンロードを行うと、ダウンロードしたファイルの拡張子が変更され、以下に記述されていることが出来なくなります。

R のメニューの “パッケージとデータ” の “パッケージインストーラ”

[*4] http://cran.md.tsukuba.ac.jp/



を選択し、“CRAN(バイナリ)” の所で



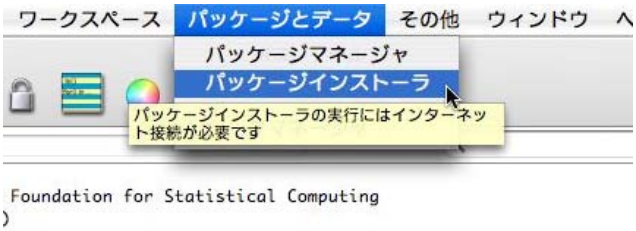
“このコンピュータ上のパッケージ” を選択し、

“インストール/アップデータ” をクリックし、ダウンロードしたファイルまたは外部記憶装置にあるファイルを選択します。

上の例は、“car_1.0-17.tgz” をインストールしています。インストール終了後、コンソールには成功したというメッセージが表示されます。

CRAN からの直接インストール

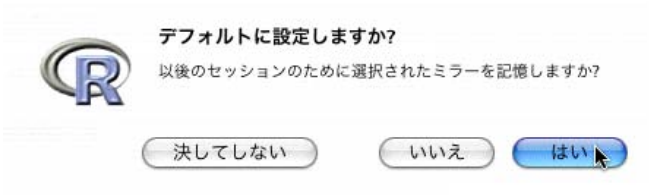
Internet に接続しており、利用しているパソコンのみにパッケージをインストールすることを考える[*5]。メニューの “パッケージとデータ” から



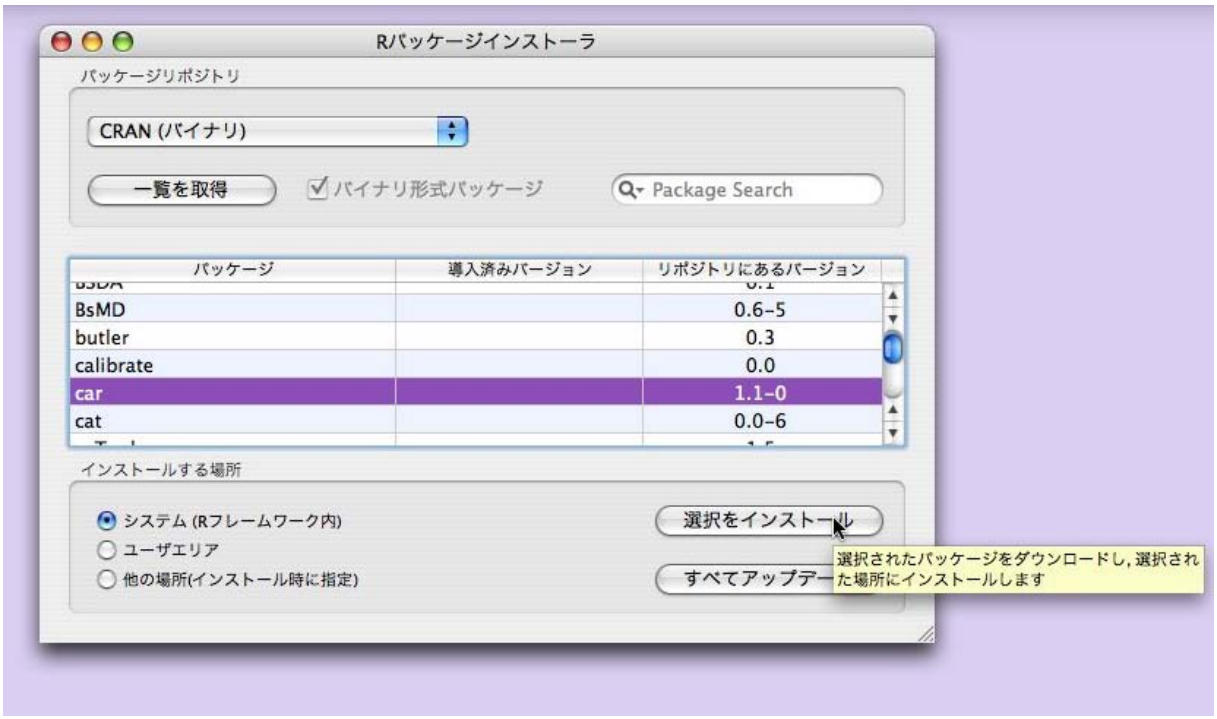
“パッケージインストーラ” を選択し、“一覧を取得”

をクリックし、ミラーサーバーとして、筑波大学

[*5] 滋賀大経済からの場合は `Sys.setenv("http_proxy"="http://cachesrv.biwako.shiga-u.ac.jp:8080/")` という一行をファイル `Rprofile` に加える必要があります。またはパッケージをインストールしたいときに `Sys.setenv("http_proxy"="http://cachesrv.biwako.shiga-u.ac.jp:8080/")` をコンソールに入力します。



を指定します。このミラーサーバーの選択・変更は R の環境設定から “起動”-“デフォルトの CRAN ミラー” を用いて行うことが出来ます。必要なパッケージを選択し、

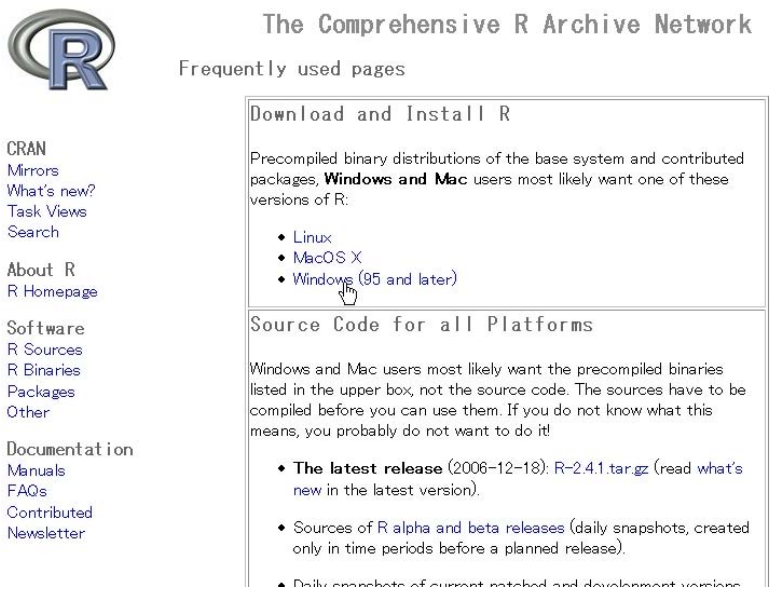


“インストール/アップデート” をクリックします。tgz ファイルをインストールしたときと同じように正しくインストールできたとき、成功したというメッセージが表示されます。

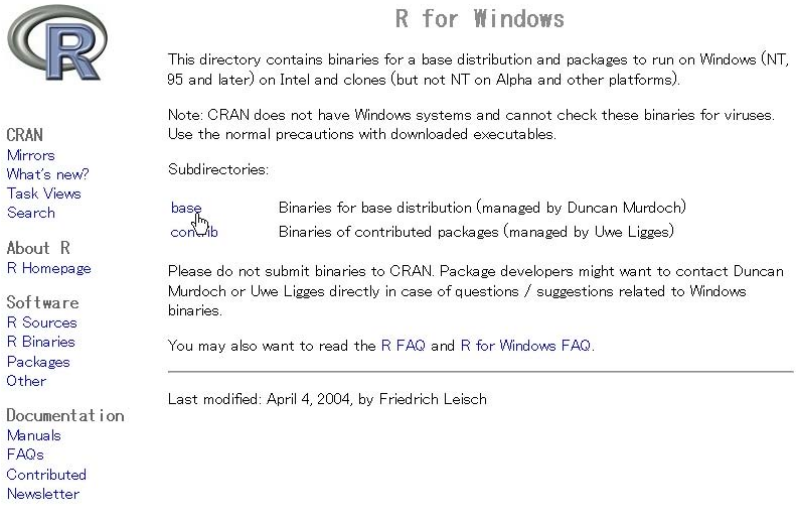
C.2　Windows 版

C.2.1　R のインストール

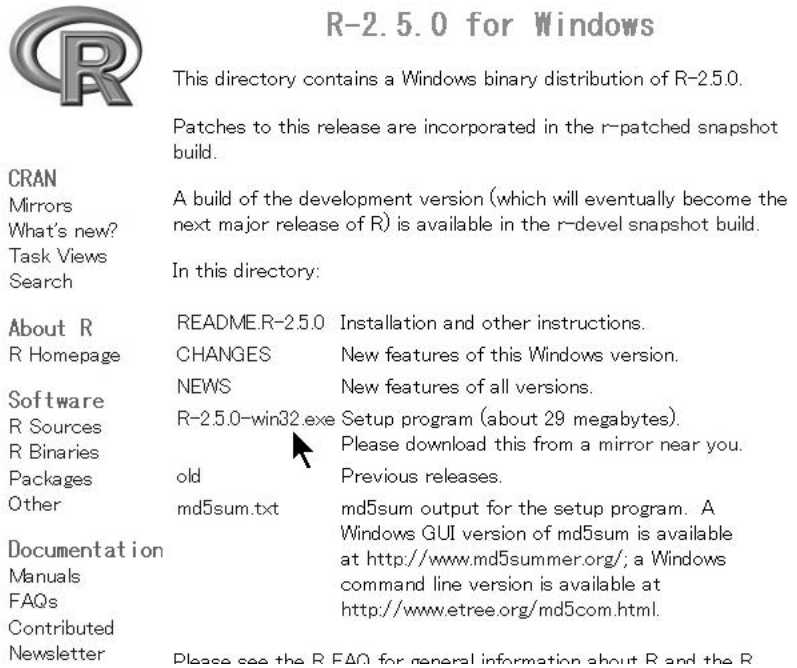
日本のミラーサーバー、例えば、筑波大学サーバーのページ



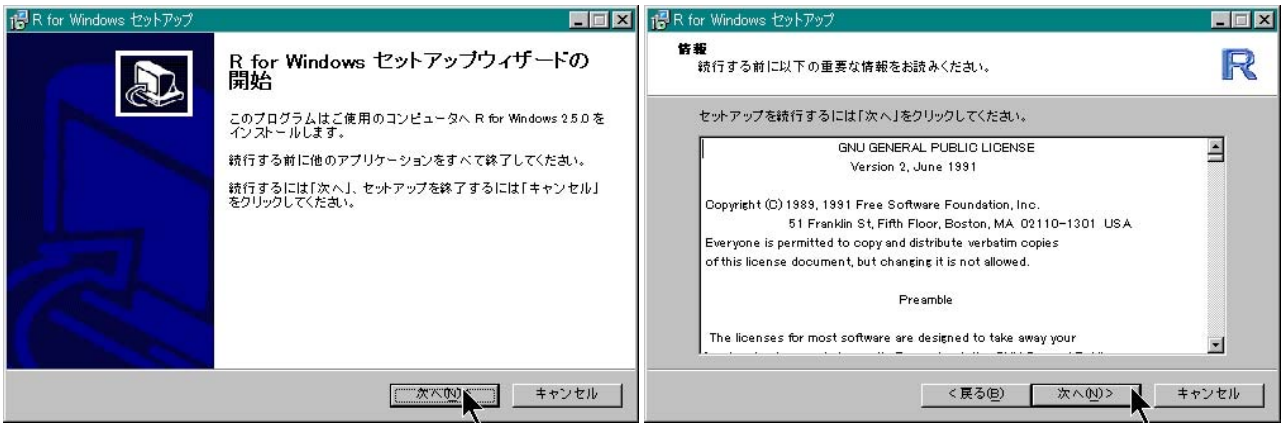
の Windows (95 and later) をクリックし、



base をクリックし、

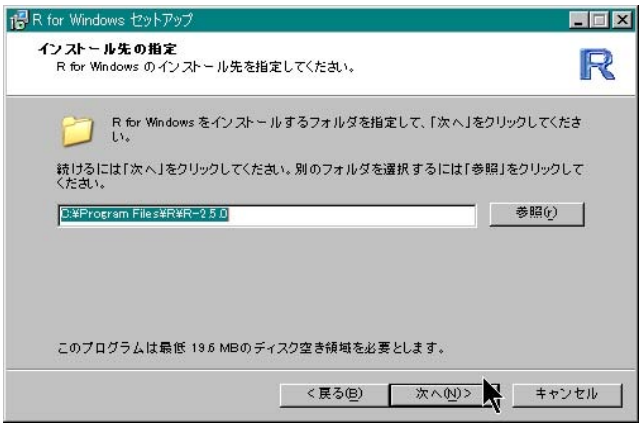


“R-2.5.0-win32.exe” をクリックすることでダウンロードが始まり、ハードディスクの適当な所に保存されます。
このファイルをクリックすることで、ソフトのインストールが始まります[*6]。
始めに、インストール中に使用する言語を選択し、

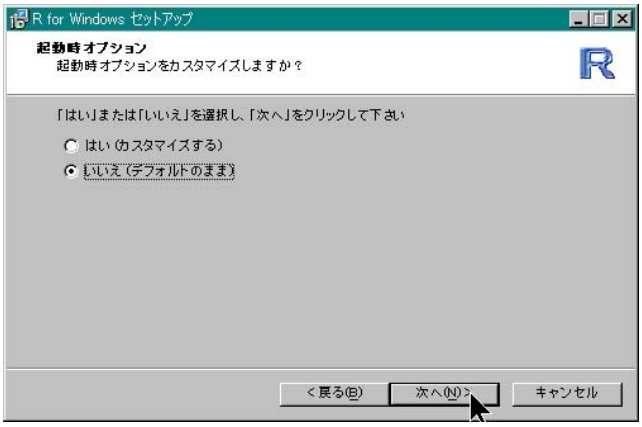


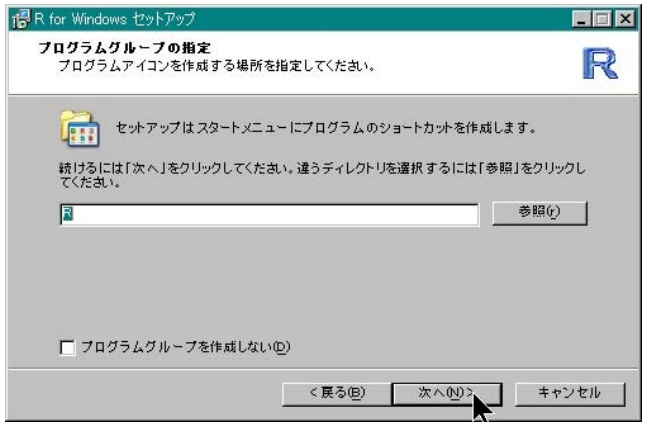
“次へ (N)>” をクリックし、使用許諾を読み、

[*6] インストール時は R のバージョンが 2.5.0 で、WindowsXP です。Windows のデザインをクラシックスタイルに変更しています。

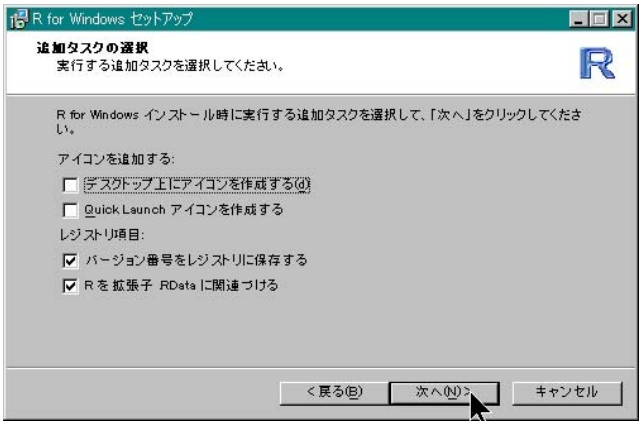


インストール先を設定し (デフォルトのままでよい)、“次へ (N)>” をクリックし、



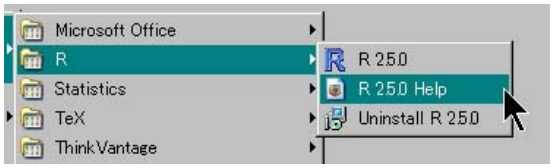


起動時オプションは “いいえ (デフォルトのまま)” を選択し、



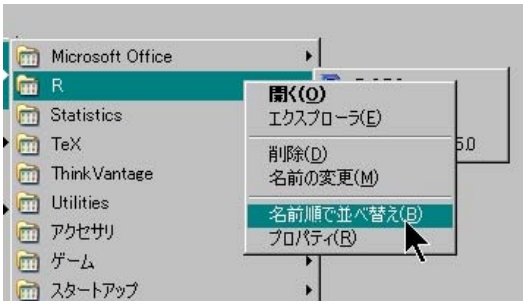
インストールが終了します。

“スタート”-“プログラム” に R というフォルダーが作成され、



その中に “R 2.5.0” というアイコンが作成されています。このアイコンまたはデスクトップの R のアイコンをクリックすることで、R が起動されます。上の図は日本語版 “R 2.5.0” をインストールしたときのものです。

スタート-プログラム内のメニューは、どの項目でも良いが、右クリックして



“名前順で並べ替え (R)” を選択すると名前順に整列し、ソフトのフォルダを見つけやすくなります。

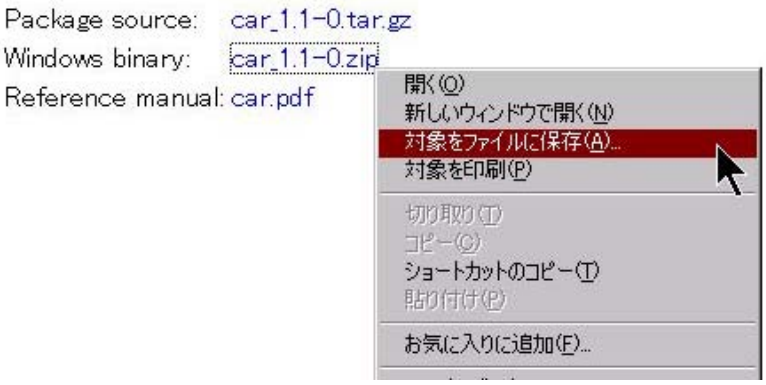
C.2.2　パッケージのインストール

HDD に zip ファイルをダウンロードした後インストール

パッケージを HDD にダウンロードし、それを R に組み込むことを説明します。これはネットに繋がっている状態のときに必要なパッケージをダウンロードし、その後他のパソコンにもインストールするときにも利用できます。CRAN ウェブページ[*7]の左フレームの “Packages” をクリックすると、右のフレームに “Contributed Packages” が現れます。

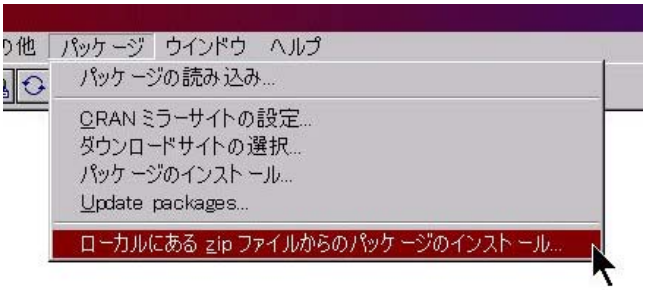
この “Contributed Packages” からダウンロードしたい Package 名 (以下の例は car の場合) を探し、Package source“car_1.1.0.zip” を右クリックし “対象をファイルに保存 (A)...” を選択し、

[*7] http://cran.md.tsukuba.ac.jp/

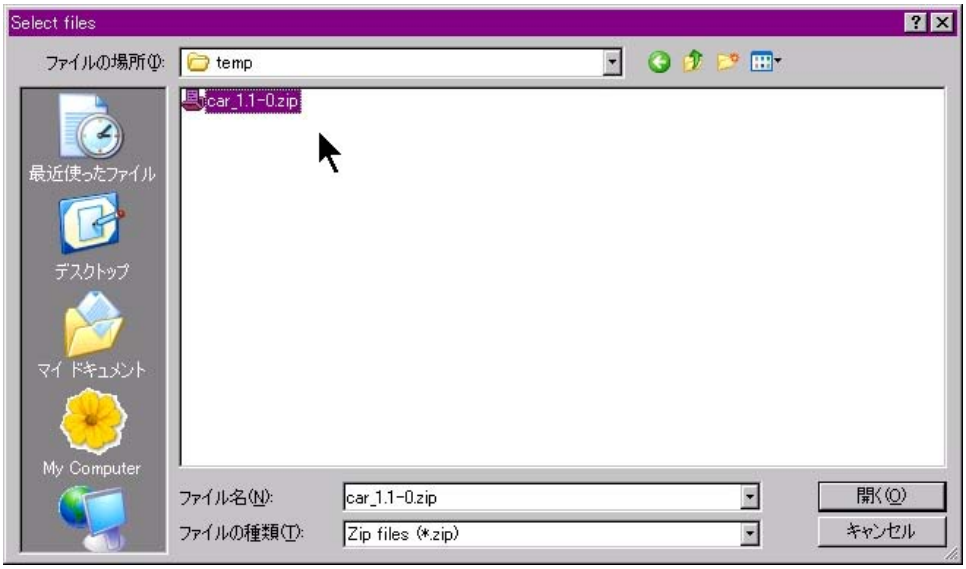


適当な所にダウンロードします。他の機械にパッケージをインストールするときは、これを USB メモリー等の外部記憶装置に保存しておきます。

R のメニューの “パッケージ” の “ローカルにある zip ファイルからのパッケージのインストール...” を選択し、



先ほど、ダウンロードしたファイルまたは外部記憶装置のファイルを指定します。



上の例は、“car_1.1-0.zip” をインストールしています。インストール終了後、コンソールには成功したというメッセージが表示されます。

CRAN からの直接インストール

Internet に接続しており、利用しているパソコンのみにパッケージをインストールすることを考える[*8]。メニューのパッケージから

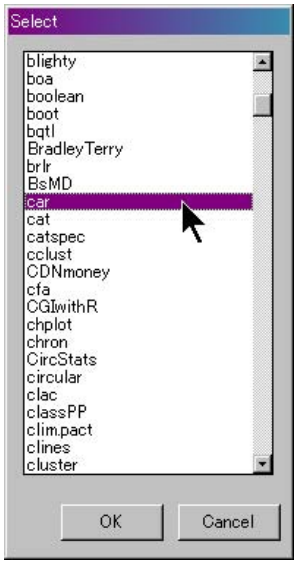
“CRAN ミラーサイトの設定...” を選択し、日本のミラーサイト

を指定します。

メニューから “パッケージのインストール...”

[*8] 滋賀大経済からの場合は `Sys.setenv("http_proxy"="http://cachesrv.biwako.shiga-u.ac.jp:8080/")` という一行をファイル `Rprofile.site` に加える必要があります。またはパッケージをインストールしたいときに `Sys.setenv("http_proxy"="http://cachesrv.biwako.shiga-u.ac.jp:8080/")` をコンソールに入力します。

を選択し、インストールしたいパッケージを選びます。



上の例では car をしていしています。zip ファイルからのインストールと同じように正しくインストールできたとき、成功したというメッセージがコンソールに表示されます。

Tips demo について

入力

```
> demo()     # demo の表示。
```

demo() とすると 13 種類のデモンストレーションの一覧が表示されます。興味深いものとしては数式の例を表示する demo(plotmath) と R を用いて描画される図の例を表示する demo(graphics) があります。

付録 D　エディタについて

D.1　エディタ

R の関数またはデータを R の console に入力し、enter キーを押すことで対話形式で結果が得られる。しかし、関数やデータをきれいな形で保存する機能は R には備わっていない。そこで、R の関数やデータをテキスト形式で保存しておくためにエディタを利用することを考える。OS に標準でインストールされているテキストエディット (Mac OS X) やメモ帳 (Windows) を利用してもよいが、検索機能、置換機能、行番号表示がない、ファイルサイズに制限がある等の理由で、フリーソフトの利用を考えます。Word 等のソフトを利用することも可能であるが、動作が速い、他の OS に持っていっても編集が出来る等の理由でエディタを用います。

エディタに入力したものを R のコンソールに Copy & Paste して作業します。これには OS によるキー割り当て

	Mac OS X	Windows
コピー	コマンド +C	Ctrl +C
貼り付け	コマンド +V	Ctrl +V
すべてを選択	コマンド +A	Ctrl +A
やり直し	コマンド +Z	Ctrl +Z
上書き保存	コマンド +S	Ctrl +S
名前を付けて保存	コマンド +Shift+S	Ctrl +Shift+S
検索	コマンド +F	Ctrl +F

C、V、A、Z、S、F 等は大文字、小文字関係なく利用できます。

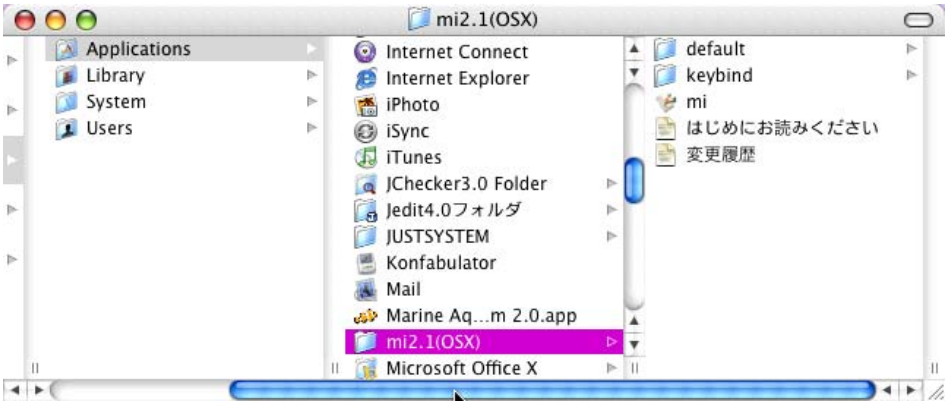
を使うと便利です。例えば、Mac OS X の場合、エディタのテキストの内容をすべて R の console に貼り付けるには、コマンドキーを押しながら A を押し、コマンドキーを押しながら C を押し、R の console をクリックして、コマンドキーを押したままで V を押せば、貼り付けが終わります。Windows の場合、コマンドキーは Ctrl キーに替わります[*1]。

D.1.1　Mac OS X の場合

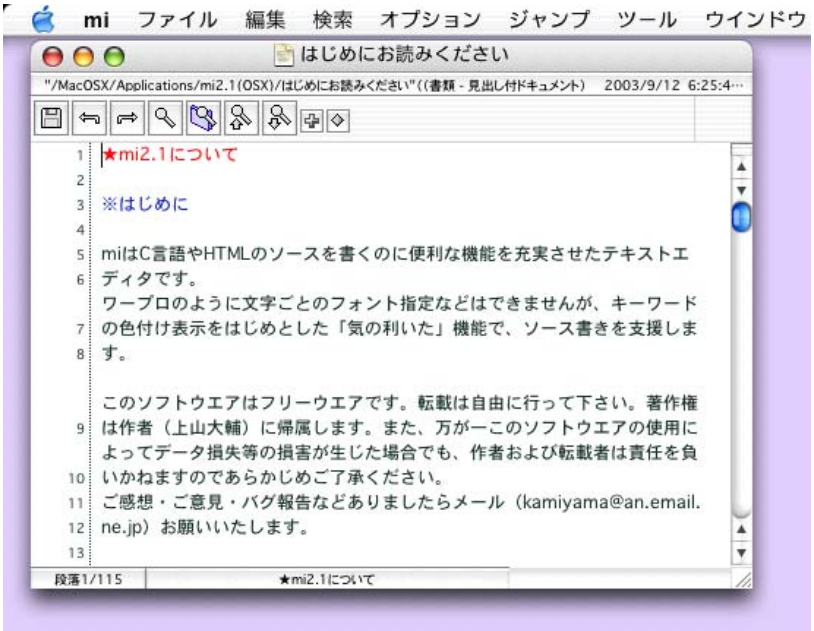
フリーソフトのミミカキエディト[*2]をインストールすることを考えます。ダウンロードのページ[*3]より、最新版をクリックし、ファイル[*4]をダウロードします。解凍が自動的に始まり、解凍が終わるとデスクトップに下のように表示されます。

[*1] Windows の一部のソフトでは、このキー割り当てを採用していないものがあります。
[*2] http://www.mimikaki.net/
[*3] http://www.mimikaki.net/download/index.html
[*4] 本稿作成時の正式版はバージョン 2.1.5 でした。

このフォルダーの中にある “mi2.1(OSX)” をアプリケーションのフォルダーに移動し、



このアプリケーションフォルダに移動した “mi2.1(OSX)” の中にある「はじめにお読み下さい」をクリックすると設定が終わり、

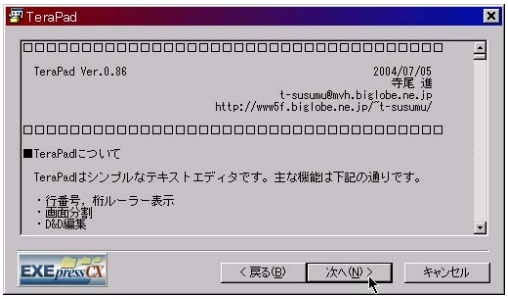


が表示されます。ダウロードしたとき出来た右図のフォルダーは不要ですので、ツールバーの矢印をクリックし削除します。“mi2.1(OSX)” の中の mi をドックに保存することで使いやすくなります。

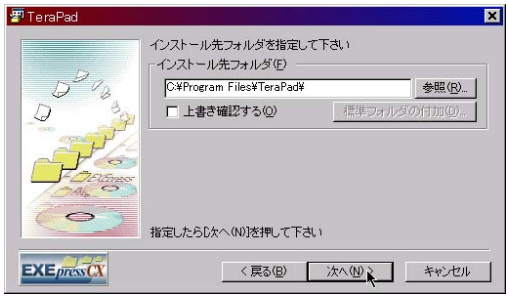
D.1.2　Windows の場合

フリーソフト TeraPad[*5]をインストールすることを考えます。ダウンロードのページ[*6]より、最新版をクリックし、ファイル[*7]をダウロードします。ダウンロードが成功すると、ファイル

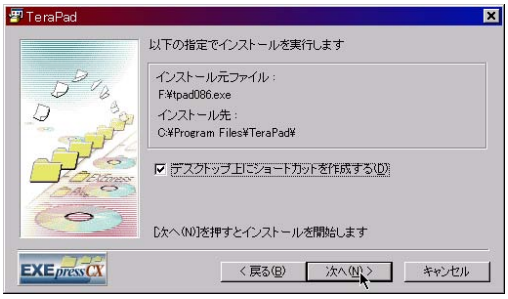
が利用しているパソコンのハードディスクに保存されます。このファイルをダブルクリックすることでインストールが始まります。Internet Explorer を利用しているときは、実行するかの問い合わせがあるので、実行を選択します。



TeraPad の保存先を指定し (デフォルトのままでよい。“上書きを確認する” にチェックマークを付けると古い TeraPad があるとき、上書き確認の確認の画面が出ます)、



初めてインストールするときは上の右図が表示されます。インストールが始まり、最後に説明書が表示されます。



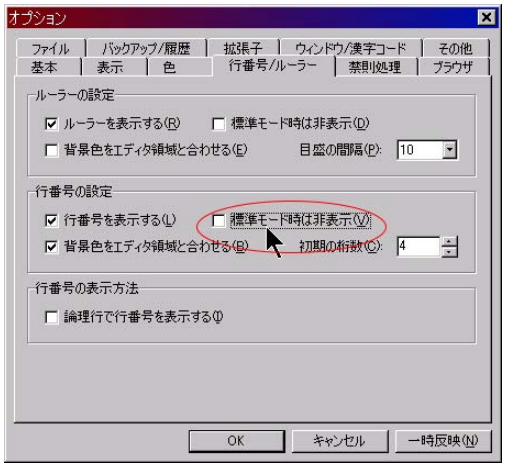
スタート-プログラムの中に TeraPad のフォルダーが作られ、TeraPad

[*5] http://www5f.biglobe.ne.jp/~t-susumu/
[*6] http://www5f.biglobe.ne.jp/~t-susumu/library/tpad.html
[*7] 本稿作成時の正式版は tpad088.exe でした。

を選択することで、TeraPad が起動します。

インストールが終わったので、ダウンロードしたファイル tpad088.exe を削除し、ゴミ箱を空にします。

TeraPad の表示-オプションで “行番号/ルーラー” のタブで “標準モード時は非表示” のチェックマークを外しておくと、テキストファイル (拡張子が txt) の時も行番号が表示されます。



Tips　ソフトについて

ソフトウェアは、店頭に製品として並んでいるものやネットからダウンロードし利用するものがあります。ネットからダウンロードしたものには、シェアウェア (shareware:代金をクレジットカード、銀行振り込み等で送金し、利用できるパスワードが送られてくる)、フリーウェア (freeware:作成者は対価を要求していないもの)、カンパウェア (ビール券、ユニセフへの送金を求められる)、メイルウェア (使っている旨をメールで伝え、感想を送る) 等がある。いずれの形態のソフトウェアも著作権が保護されているので、以下のようなことが明記されています。

1. 第三者に対し、本ソフトウェアを販売ならびに販売を目的とした宣伝、展示、使用、複製、営業等を行うこと。
2. 第三者に対し、本ソフトウェアの使用権を譲渡あるいは再許諾すること。
3. 第三者に対し、本ソフトウェアを貸与、リースもしくは担保設定すること。
4. 本ソフトウェアをリバースエンジニアリング、逆コンパイル、逆アセンブルすること、または派生商品を作成すること。
5. 本許諾書その他の付属文書を含め、本ソフトウェアの一部または全部を改変あるいは除去すること。

R は、オープンソースフリーウェアで、Free Software Foundation の GNU General Public License の条項の下で、ソースコードの形も入手できます。様々な Unix プラットフォームや、類似のシステム (FreeBSD や Linux を含む) では、ダウンロードし、コンパイルし、実行形式のファイルが得られます。また、Mac OS X や Windows でもコンパイラを用いることで実行プログラムを作成することが出来ます。

参考文献

R については

R Development Core Team (2005). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

統計学の入門書に関しては下記の文献を参考にされたい。

- 白旗 慎吾著『統計解析入門』(共立出版)
- 稲垣 宣生著『数理統計学 (改訂版) 』(裳華房)
- 栗栖 忠・濱田 年男・稲垣 宣生著『統計学の基礎』(裳華房)
- 長畑 秀和『統計学へのステップ』(共立出版)
- 長畑 秀和『多変量解析へのステップ』(共立出版)
- 杉山 高一『統計学入門』(絢文社)

R の文献に関しては下記のものが日本語としては参考になります。

- 中澤 港著『R による統計解析の基礎』(ピアソンエデュケーション)
- 間瀬 茂・神保 雅一・鎌倉 稔成・金藤 浩司著『工学のためのデータサイエンス入門』(数理工学社)
- 岡田 昌司他著『THE R BOOK』(九天社)
- 舟尾 暢男『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』(九天社)
- 荒木 孝治編著『フリーソフトウェア R による統計的品質管理入門』(日科技連出版)
- 舟尾 暢男 高浪 洋平『データ解析環境「R」—定番フリーソフトの「基本操作」から「グラフィックス」「統計解析」まで』(工学社)
- 竹内 俊彦『はじめての S-PLUS/R 言語プログラミング』(オーム社)
- 垂水 共之・飯塚 誠也『R/S-PLUS による統計解析入門』(共立出版)
- 椿広計『ビジネスへの統計モデルアプローチ』(朝倉書店)
- 赤間 世紀・山口 善博『R による統計入門』(技報堂出版)

ウェブサイトとして、

- http://www.okada.jp.org/RWiki/：RjpWiki
- http://aoki2.si.gunma-u.ac.jp/R/：R による統計処理/群馬大学青木繁伸先生

があり、有益な情報が得られます[*1]。

また、他の言語を用いたパソコンでの統計処理やノンパラメトリック統計学、医療統計に関しては下記のものを参考に挙げておく。

- 丘本 正『パソコン統計学入門』(共立出版)
- 杉山 高一・牛沢 賢二著『パソコンによる統計解析』(朝倉書店)
- ハエック (著) 丘本 正 (訳)『ノンパラメトリック統計学』(日科技連出版社)
- 鈴木 治郎 (著)『Excel で実験する統計学』(アジソンウェスレイ)
- J.L. フライス (著) 森川敏彦・熊澤吉起監修訳『臨床試験のデザインと解析』(アーム株式会社)

[*1] これらのウェブサイトを含めたものは http://www.biwako.shiga-u.ac.jp/sensei/kumazawa/R/ においてリンクが張られています。

ギリシャ文字

表.1　ギリシャ文字一覧

読み方	大文字	小文字	読み方	大文字	小文字
alpha	A	α	beta	B	β
gamma	Γ	γ	delta	Δ	δ
epsilon	E	ϵ, ε	zeta	Z	ζ
eta	H	η	theta	Θ	θ, ϑ
iota	I	ι	kappa	K	κ
lambda	Λ	λ	mu	M	μ
nu	N	ν	omicron	O	o
xi	Ξ	ξ	pi	Π	π, ϖ
rho	P	ρ, ϱ	sigma	Σ	σ, ς
tau	T	τ	upsilon	Υ	υ
phi	Φ	ϕ, φ	chi	X	χ
psi	Ψ	ψ	omega	Ω	ω

索引

R の関数・データ索引	……	112
一般・統計用語索引	……	114
英語索引	……	116

Tips 索引

`help` について— 1	……	2
乱数について	……	4
`library` と `data` について	……	6
`example` について	……	13
`help` について— 2	……	14
`option` について	……	22
履歴について	……	31
WebBrowser の操作について	……	83
HDD 内の検索ソフトについて	……	85
`demo` について	……	104
ソフトについて	……	108

R の関数・データ索引

記号

: 18
; 21

A

abline 11, 69, 70
abs 27
as.vector 26
ask 13
attach 21, 22

B

barplot 6, 7
boxplot 10, 26, 31, 46, 58, 90

C

c 16, 17, 19, 31
car 4, 18
cars 3, 11, 28, 64, 65, 67, 70
cause.and.effect 72
cbind 57
chisq.test 54–57
co2 4, 19
colnames 17
cor 28, 64, 67
cor.test 66, 67
cos 21
cov 28

D

data 1, 6, 61
data.frame 17, 31
Davis 18
decompose 5, 19
demo 104
density 90
diff 27

E

e1071 29, 30
ecdf 62
Ericksen 4
example 13, 22
exp 21, 64, 67

F

fivenum 25, 31
fix 18

H

help 1, 2, 6, 7, 10
hist 3, 7, 26, 90
history 31

I

identify 18
IQR 27
iris 10, 11, 31, 33, 34, 36, 37, 39, 41, 53, 61, 64, 66

K

ks.test 60, 61
e1071 30

L

length 23, 24, 27, 29–31, 36, 37, 64, 67
library 4, 6, 72
lines 62
list 21
lm 11, 69, 70
log 21, 67
ls 21

M

margin.table 17, 18, 57
matrix 17, 57
max 25
mean 23, 27, 29–31 , 34, 36, 37, 46, 47, 61
median 24
mfrow 7, 9, 10, 58, 90
min 24
month.name 16
mosaicplot 12, 17

N

NA 15
names 4, 8

O

options 22
outer 64

P

par 7, 9, 10, 13, 58, 90
pareto.chart 8
persp 64
pie 7
plot 3, 5, 10, 11, 13, 19, 62, 64, 69, 70
pnorm 60
prod 23

Q

qcc 8, 22, 72, 74–79
qchisq 39, 41
qnorm 34, 88, 89
qqline 9, 58, 90
qqnorm 9, 58, 90

R

range 27
read.csv 20
rep 31
rm 21
rnorm 4, 6, 9, 10, 26, 58
rownames 17
rug 58, 90
runif 3, 4, 9, 14, 58

S

sample 62, 66
sd 27, 29, 30, 61
search 6
seq 18, 64
set.seed 3, 6, 14
setwd 20
shapiro.test 58, 61

sin 21
sink 20
skew 29, 30
e1071 29
sleep 1, 2, 53
sort 31
sqrt 21, 27, 29, 30, 34, 36, 37, 47, 64
stem 6
sum 24, 27, 31, 39, 41
summary 12, 25, 26, 31

T

t.test 36, 37, 43, 47–50
t.var 47
table 62, 66
tan 21
Titanic 3, 12, 56, 57
title 90
ToothGrowth 45–52
ts 19

V

VADeaths 6, 7
var 26, 36, 37
var.test 51, 52

一般・統計用語索引

C

Copernic Desktop Search 85
csv ファイル 20

E

Excel 29, 30

G

goo 81
google 81

I

iris 60

K

Kolmogorov-Smirnov 検定統計量 59, 60

P

Personal Dictionary for Win32 81

S

SAS 20
SPSS 20

T

TeraPad 107
TEX 4

W

Wikipedia 80

Y

Yahoo! 81

あ

R 管理図 75

い

一様乱数 3, 14
インターネット検索 81

う

ウィキペディア 80
ウェルチの検定 49
上側ヒンジ 23, 25, 31

え

英辞郎 81
英辞郎ビューアー 81
S 管理図 74
x 管理図 76
$\bar{x}$ 管理図 74
エディタ 105
F 検定 51
F 分布 50, 51
円グラフ 7

お

大きさ 32, 44
重み付け平均 24

か

回帰係数 68
カイ二乗検定 39, 41, 54–57
カイ二乗分布 38, 40, 54–57
拡張子 84, 85
確率密度プロット 90
片側検定 32, 35, 37, 42, 89
間隔尺度 1
ガンマ関数 36
管理図 72
　c 管理図 78
　p 管理図 76
　pn 管理図 77
　R 管理図 75
　S 管理図 74
　u 管理図 78
　x 管理図 76
　$\bar{x}$ 管理図 74

き

幾何平均 23, 31
調和平均 23, 24, 31
棄却領域 32, 86
既知 32, 44
帰無仮説 32, 35, 37, 42, 86, 89
共分散 23, 28, 31
行和 17

け

経験分布関数 59
決定係数 70, 71
検出力 87
検定 32, 44, 86

こ

Kolmogorov-Smirnov 検定 59

さ

最小二乗法 69
最小値 23, 24, 31
最大値 23, 25, 31
最頻値 23, 24, 31
算術平均 23, 31
散布図 3, 11, 28, 63

し

c 管理図 78
シェアウエア 108
時系列データ 4
四則演算 21
下側ヒンジ 23, 25, 31
質的データ 1
Gini の係数 30
四分位範囲 23, 27, 31
Shapiro-Wilk の正規性検定 58
重回帰 68
従属変数 68
自由度 35, 38, 40, 42, 47, 49, 54–57, 65
樹葉図 6
シュワルツの不等式 63
順序尺度 1

順序統計量 24

す

スタージェスの公式 8

せ

正規確率プロット 9, 13, 90
正規分布 9, 29, 32, 33, 40, 42, 44, 58, 88
正規密度関数 13, 33, 64, 65
積率 29
z 変換 67
説明変数 68
尖度 23, 29, 31

そ

相関係数 23, 28, 31, 63
ソフトウエア 108

た

第一種の過誤 87
第二種の過誤 87
第 1 四分位点 10, 23, 25, 31
第 3 四分位点 10, 23, 25, 31
対立仮説 32, 35, 37, 42, 86, 89
単回帰 68

ち

中央値 10, 23, 24, 31

つ

対になった標本に関する検定 42
積み上げ棒グラフ 7

て

t-検定 36, 37
t 分布 35, 42, 47, 49, 65

と

特性要因図 72
独立変数 68
度数分布図 6

に

二変量正規分布 64, 65

は

箱ひげ図 10, 13, 90
外れ値 10
パッケージ 4, 94, 101
パラメータ 32
パレート図 8
範囲 23, 27, 31

ひ

pn 管理図 77
p 管理図 76
p-値 36, 37, 43
p 値 47, 54, 57, 59–61, 66, 70, 71
ヒストグラム 7, 8, 13, 90
標準偏差 23, 27, 31
標本 32
標本分散 26, 35, 40
標本平均 23
比例尺度 1
ヒンジ散布度 23, 27, 31
品質管理 7

ふ

プールする 47
不偏分散 26, 35
フリーウエア 4, 108
分割表 3, 55
分散 23, 25, 31, 32, 38, 40, 45, 50

へ

平均の差 45
平均偏差 27
ベクトル 16
偏差 25, 26
変動係数 30

ほ

母集団 32, 44, 58
母数 32
母分散 32, 35, 37, 40

み

未知 32, 35, 37, 40, 44
密度関数 13
ミミカキエディト 105

む

無相関の検定 65

め

名義尺度 1

も

モーメント 29
目的変数 68
モザイクプロット 12

ゆ

有意水準 32, 35, 37, 45, 86, 87
u 管理図 78

ら

乱数 4

り

離散型データ 1
両側検定 32, 35, 37, 42, 89
量的データ 1

れ

列和 17
連続型データ 1

わ

歪度 23, 29, 31

英語索引

A

alternative（対立仮説）..... 32, 35, 37, 42, 86, 89
arithmetic mean（算術平均）..... 23, 31

B

bar chart（積み上げ棒グラフ）..... 7
boxplot（箱ひげ図）..... 10, 90

C

cause and effect diagram（特性要因図）..... 72
c control chart（*c* 管理図）..... 78
χ^2 distribution（カイ二乗分布）..... 38, 40, 54–57
χ^2 test（カイ二乗検定）..... 39, 41, 54–57
coefficient of determination（決定係数）..... 70, 71
coefficient of variation（変動係数）..... 30
comma separated value file（csv ファイル）..... 20
contingency table（分割表）..... 3, 55
continuous data（連続型データ）..... 1
control chart（管理図）
  c control chart（*c* 管理図）..... 78
  p control chart（*p* 管理図）..... 76
  pn control chart（*pn* 管理図）..... 77
  R control chart（*R* 管理図）..... 75
  S control chart（*S* 管理図）..... 74
  u control chart（*u* 管理図）..... 78
  $\bar{x}$ control chart（$\bar{x}$ 管理図）..... 74
  x control chart（*x* 管理図）..... 76
control chart（管理図）..... 72
correlation coefficient（相関係数）..... 23, 28, 31, 63
covariance（共分散）..... 23, 28, 31
CRAN（Comprehensive R Archive Network）..... 91
criterion variable（目的変数）..... 68

D

degree of freedom（自由度）..... 35, 38, 40, 42, 47, 49, 54–57, 65
density plot（確率密度プロット）..... 90
dependent variable（従属変数）..... 68
deviation（偏差）..... 25, 26
discrete data（離散型データ）..... 1

E

empirical cumulative distribution function（経験累積分布関数）..... 59
explanatory variable（説明変数）..... 68

F

F distribution（*F* 分布）..... 50, 51
F test（*F* 検定）..... 51
first quantile（第 1 四分位点）..... 23, 25, 31
frequency histogram（度数分布図）..... 6

G

gamma function（ガンマ関数）..... 36
geometric mean（幾何平均）..... 23, 31
Gini coefficient（Gini の係数）..... 30

H

harmonic mean（調和平均）..... 23, 24, 31
hinge spread（ヒンジ散布度）..... 23, 27, 31
histogram（ヒストグラム）..... 7, 90

I

independent variable（独立変数）..... 68
interquantile range（4 分点範囲）..... 23, 27, 31
interval scale（間隔尺度）..... 1

K

known（既知）..... 32, 44
Kolmogorov-Smirnov test statistic（Kolmogorov-Smirnov 検定統計量）.. 59, 60
Kolmogorov-Smirnov test（Kolmogorov-Smirnov 検定）..... 59
kurtosis（尖度）..... 23, 29, 31

L

least square method（最小二乗法）..... 69
lower hinge（下側ヒンジ）..... 23, 25, 31

M

maximum（最大値）..... 23, 25, 31
mean deviation（平均偏差）..... 27
median（中央値）..... 10, 23, 24, 31
minimum（最小値）..... 23, 24, 31
mode（最頻値）..... 23, 24, 31
moment（モーメント）..... 29
mosaic plot（モザイクプロット）..... 12
multiple regression（重回帰）..... 68

N

nominal scale（名義尺度）..... 1
normal distribution（正規分布）..... 9, 29, 32, 33, 40, 42, 44, 58, 88
normal quantile-quantile plot（正規確率プロット）..... 9, 90
null hypothesis（帰無仮説）..... 32, 35, 37, 42, 86, 89

O

one-sided test（片側側検定）..... 32, 35, 37, 42, 89
order statistics（順序統計量）..... 24
ordinal scale（順序尺度）..... 1

P

p-value（*p*-値）..... 36, 37, 43, 47, 54, 57, 59–61, 66, 70, 71
parameter（母数）..... 32
Pareto chart（パレート図）..... 8
p control chart（*p* 管理図）..... 76
pie chart（円グラフ）..... 7
pn control chart（*pn* 管理図）..... 77
population（母集団）..... 32, 44, 58
power（検出力）..... 87

Q

qualitative data（質的データ）..... 1
quantitative data（量的データ）..... 1

R

range（範囲）..... 23, 27, 31
ratio scale（比例尺度）..... 1
R control chart（*R* 管理図）..... 75
regression coefficient（回帰係数）..... 68
rejection region（棄却領域）..... 32, 86

S

sample（標本）..... 32
sample mean（標本平均）..... 23
sample variance（標本分散）..... 26, 35, 40
scatter diagram（散布図）..... 11, 63
Schwarz inequality（シュワルツの不等式）..... 63
S control chart（*S* 管理図）..... 74

Shapiro-Wilk test for normality（Shapiro-Wilk の正規性検定）.....58
significance level（有意水準）.....32, 35, 37, 45, 86, 87
simple regression（単回帰）.....68
size n（大きさ n）.....32, 44
skew（歪度）.....23, 29, 31
standard variance（標準偏差）.....23, 27, 31
stem and leaf plot（樹葉図）.....6
Sturges formula（スタージェスの公式）.....8

T

t distribution（t 分布）.....35, 42, 47, 49, 65
t-test（t 検定）.....36, 37
third quantile（第 3 四分位点）.....23, 25, 31
two-sided test（両側検定）.....32, 35, 37, 42, 89
Type I Error（第一種の過誤）.....87
Type II Error（第二種の過誤）.....87

U

u control chart（u 管理図）.....78
unbiased variance（不偏分散）.....26, 35
unknown（未知）.....32, 35, 37, 40, 44
upper hinge（上側ヒンジ）.....23, 25, 31

V

variance（分散）.....23, 25, 31, 32, 38, 40, 45, 50

W

weighted mean（重み付け平均）.....24
Welch test（Welch の検定）.....49

X

$\bar{x}$ control chart（$\bar{x}$ 管理図）.....74
x control chart（x 管理図）.....76

Z

z-transformation of the correlation（z 変換）.....67