

統計データ解析

R 言語の基礎

吉田朋広 (東京大学)
小池祐太 (東京大学)
村田 昇 (早稲田大学・東京大学)

version: 2020 年 3 月 19 日

東京大学大学院数理科学研究科
統計データ解析教育研究グループ[°]

目次

1 R の基本的な操作	1
1.1 はじめに	1
1.1.1 R 言語	1
1.1.2 起動と終了	2
1.1.3 ヘルプ機能	3
1.1.4 パッケージ管理	3
1.2 基本的な使い方	4
1.2.1 式の入力	4
1.2.2 数の扱い	5
1.2.3 変数への代入	6
1.3 データ構造	6
1.3.1 ベクトル	7
1.3.2 行列	9
1.3.3 リスト	10
1.3.4 データフレーム	12
1.4 補遺	13
1.4.1 参考文献	13
1.4.2 亂数の生成	14
2 ベクトル・行列の演算と関数	15
2.1 ベクトルの計算	15
2.1.1 和	15
2.1.2 積	16
2.1.3 初等関数の適用	16
2.2 行列とその演算	17
2.2.1 和	17
2.2.2 積	18
2.2.3 初等関数の適用	19
2.2.4 行列式とトレース	19
2.2.5 逆行列	20
2.3 ベクトルと行列の計算	21
2.3.1 行列とベクトルの積	21
2.3.2 連立方程式	21
2.4 関数	22
2.4.1 制御文	22
2.4.2 関数の定義	23
2.5 補遺	24
2.5.1 参考文献	24
2.5.2 ベクトルのノルム	24
2.5.3 行列のノルム	25
2.5.4 一般化逆行列	26
2.5.5 固有値と固有ベクトル	27
2.5.6 特異値分解	27
3 データの加工・整理と入出力	29
3.1 データの形式	29
3.1.1 値の型	29

3.1.2 ベクトル	30
3.1.3 行列	30
3.1.4 配列	31
3.1.5 データフレーム	32
3.2 データの抽出	34
3.3 ファイルを用いたデータの読み書き	38
3.3.1 作業ディレクトリの確認と変更	39
3.3.2 CSV 形式の操作	39
3.3.3 RData 形式の操作	41
3.4 データの整理	42
3.5 補遺	45
3.5.1 参考文献	45
3.5.2 パッケージ dplyr によるデータの操作	45
3.5.3 条件を指定した行の選択	46
3.5.4 列の値による行の並べ替え	46
3.5.5 列の選択	47
3.5.6 値の整理	48
3.5.7 列の追加	49
3.5.8 データフレームの集計	49
3.5.9 行のリサンプリング	50
3.5.10 データフレームのグループ化	50
3.5.11 その他	51
4 データのプロット	53
4.1 基本的な描画	53
4.2 ヒストグラム	56
4.3 箱ひげ図	57
4.4 棒グラフ	58
4.5 円グラフ	59
4.6 散布図行列	60
4.7 3次元のグラフ	61
4.8 プロット環境の設定	62
4.9 補遺	63
4.9.1 参考文献	63
4.9.2 パッケージ ggplot2 の利用	63
4.9.3 基本的な文法	64
4.9.4 散布図	64
4.9.5 曲線あてはめ	65
4.9.6 ヒストグラム	66
4.9.7 箱ひげ図	67
4.9.8 折れ線グラフ	68
4.9.9 棒グラフ	69
5 モンテカルロ法	71
5.1 亂数	71
5.2 数値シミュレーション	73
5.2.1 コイン投げの賭け	73
5.2.2 Buffon の針	74
5.2.3 Monty Hall 問題	75
5.2.4 St Petersburg のパラドックス	77
5.2.5 秘書問題	79

5.3 補遺	82
5.3.1 參考文獻	82

R の基本的な操作

1.1 はじめに

まず、はじめに R の概要について述べる。

1.1.1 R 言語

R(または R 言語と呼ばれる) は統計計算のための言語と環境の総称であり、オープンソース・フリーソフトウェア (open source, free software) である。利用の規約は GNU General Public License (GPL) に従うので、その内容について詳しく知りたい場合は

<https://www.gnu.org/licenses/gpl-3.0.en.html>

(英語)

<https://www.gnu.org/licenses/gpl-3.0.ja.html>

(日本語)

を参照して欲しい。

また、多くの人により開発されている多数のパッケージ (package) によって、様々な機能 (パッケージは関数やデータの集合体と考えればよい) を追加することができる。R の本体、およびパッケージは、開発プロジェクトのサイト R Project (The R Project for Statistical Computing)

<https://www.r-project.org/>

のメニューにある CRAN (The Comprehensive R Archive Network) の中にあるミラーサイト (mirror site; 日本国内にもある) からダウンロードすることができる。R の本体は OS (Operating System; Linux, MacOS, Windows) 別に異なる配布物として公開されており、それぞれの OS に適切な方法で簡単にインストールすることができる。また、パッケージは R の中に用意された関数や GUI (Graphical User Interface; 画面上のグラフィックスとマウスなどを用いて直感的な操作を提供するユーザインターフェース) を用いてインストールすることができる。

R Project で公開されている R 本体には Windows や MacOS の場合は専用の GUI が用意されているが、UNIX 系 OS の場合はターミナル (シェル) から起動する必要がある。このため OS によって若干操作性が異なるという問題があるが、UNIX も含め様々な OSにおいて同様に利用することができる RStudio という統合開発環境 (integrated development environment; IDE) が RStudio 社により開発され公開されている。

<https://www.rstudio.com/>

講義では OS による操作の違いをできるだけ少なくするために、RStudio を用いて説明を行う。

演習 1.1. R と RStudio を自身の PC にインストールしてみよう。

<https://www.r-project.org/> (The R Project)
<https://www.rstudio.com/> (RStudio, inc)

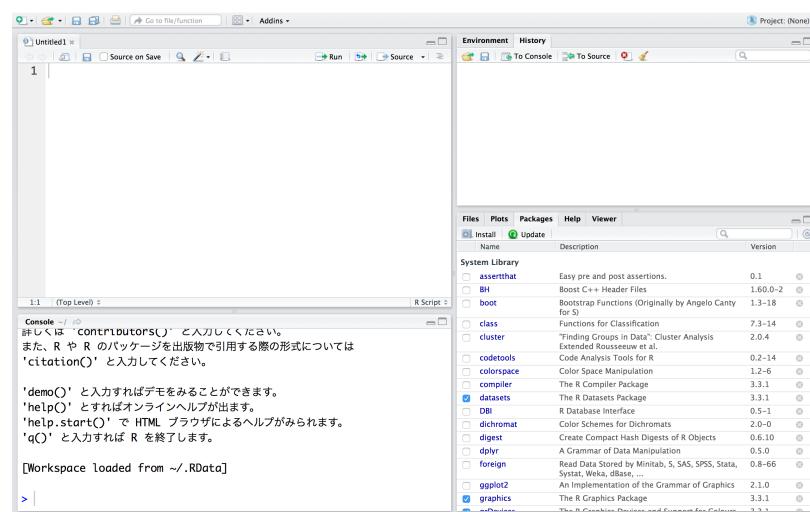
例えば以下のサイトがインストールの参考になる。

<http://www.okadajp.org/RWiki/?R%20のインストール>
<http://aoki2.si.gunma-u.ac.jp/R/begin.html>

1.1.2 起動と終了

RStudio を起動すると、標準では図 1.1 のような 4 ペイン (pane) のウインドウが立ち上がる。左上がエディタ、左下がコンソール、右上が変数や履歴、右下がグラフィクスやヘルプなどを表示するペインとなる。

図 1.1: RStudio の起動画面。



起動後、コンソールは以下のようなメッセージを表示し、最後に入力を促すプロンプトである ‘>’ 記号を表示して入力待ちの状態となる。

```
R version 3.6.2 (2019-12-12) -- "Dark and Stormy Night"
Copyright (C) 2019 The R Foundation for Statistical Computing
Platform: x86_64-apple-darwin15.6.0 (64-bit)

R は、自由なソフトウェアであり、「完全に無保証」です。
一定の条件に従えば、自由にこれを再配布することができます。
配布条件の詳細に関しては、'license()' あるいは 'licence()' と入力
してください。

R は多くの貢献者による共同プロジェクトです。
詳しく述べは 'contributors()' と入力してください。
また、R や R のパッケージを出版物で引用する際の形式については
'citation()' と入力してください。

'demo()' と入力すればデモをみることができます。
'help()' とすればオンラインヘルプが出ます。
'help.start()' で HTML ブラウザによるヘルプがみられます。
'q()' と入力すれば R を終了します。

>
```

例えばここで終了を指示する `q()` を入力すれば、R は終了する。

終了時にメッセージが表示される場合があるが、これに対して“y(yes)”を入力すると、それまでに定義された変数や関数およびコマンドの履歴(ヒストリ)が保存され、次回起動時に自動的に読み込まれる。“n(no)”になるとこのセッションで更新した内容は残らない。また、終了することを中止して計算を続ける場合は“c(cancel)”を入力する。

なお、入力文字列において‘#’以降は無視されるので、以降の実行例においては#を用いて必要なコメントを記載していく。

1.1.3 ヘルプ機能

Rにはオンラインのヘルプ機能が備えられていて、コンソール(左下のペイン)から関数 `help()` に関数名を、関数 `help.search()` には検索したいキーワードを渡すことによって利用することができる。なお、以下はあくまで一つの出力例であり、環境(Rのバージョンやインストールされているパッケージなど)によって出力が異なる場合があることに注意して欲しい。

```
> ### 関数 help の使い方
> help(sin) # 三角関数のヘルプを見る
> ?log # ? は help() の代替
> ### 関数 help.search の使い方
> help.search("histogram") # ヒストグラム関連の情報を探す
> ??random # ?? は help.search() の代替
```

Rscript: `basic-help.r`

最初の例は `sin` 関数を調べたもので、右下のペインにヘルプの内容が出る。左上の“Trig”は見出しで、この内容が“trigonometric functions”に関するヘルプであることを表わしている。また中央上の“package:base”は“base”というパッケージ内の関数であることを示している。

二番目の例は“histogram”に関する事項を検索したもので、例えば“graphics::hist”は“graphics”というパッケージ内にある“hist”という関数がヒストグラムの作成に関連することを示している。

なお、上記の例で出てきた“base”や“graphics”は指定しなくとも標準で読み込まれるパッケージである。読み込まれているパッケージを確認する方法は次節を参照して欲しい。

GUIを用いる場合は右下のペインの“Help”タブ(tab)を利用する。関数名またはキーワードを入力して必要な情報を検索することができる。

Rの本体、あるいはパッケージに関するドキュメント(マニュアル)は開発プロジェクトのサイト CRAN にあるが、使い方を含め有用な情報を解説するサイトとして

<http://www.okada.jp.org/RWiki/>
<http://aoki2.si.gunma-u.ac.jp/R/>

など数多くあるので、これらも合わせて参照して欲しい。

1.1.4 パッケージ管理

CRAN では 2018 年 4 月 3 日現在、12368 を越えるパッケージが公開されている。

右下のペインには“Packages”タブがあり、GUIを用いてパッケージ管理を行うことができる。必要な機能を持つパッケージ名を調べておけば、“Packages”タブの中の“Install”から新規にパッケージをインストールすることができる。また“Update”を選ぶとインストール済のパッケージの更新を行うことができる。なお、標準でいくつかのパッケージは自動的に読み込まれており、既に読み込まれたパッケージは“Packages”タブで確認することができる。

関数 `install.packages()` を用いればコンソールから直接インストールすることができる。以下は一つの出力例であり、環境によっては異なる場合もあることに注意して欲しい。

```
> install.packages("ggplot2",repos='https://cran.istm.ac.jp/')

ダウンロードされたパッケージは、以下にあります
  /var/folders/abc/xyz downloaded_packages
>
```

パッケージを取り扱う関数についての更に詳しい情報は

```
help("install.packages"),
help("update.packages"),
または
help("INSTALL")
```

などを利用して調べて欲しい。

1.2 基本的な使い方

1.2.1 式の入力

四則演算や一般的な関数はC言語などの計算機言語とほぼ同じ名称で使うことができ、直感に沿った文法で計算を実行することができる。

Rscript: `basic-calc.r`

```
> (1 + 3) * (2 + 4) / 6 # 四則演算
[1] 4
> 1.8 + 5 - 0.04 + 8.2 / 3 # 計算順に注意
[1] 9.493333
> pi # π（パイ）は定義されている
[1] 3.141593
> print(pi,digits=22) # 桁数を変更して表示
[1] 3.141592653589793115998
> sqrt(2) # 平方根
[1] 1.414214
> 8^(1/3) # 置乗
[1] 2
> exp(10) # 指数関数
```

```
[1] 22026.47
> exp(1) # 自然対数の底
[1] 2.718282
> log(10) # 対数関数 (log, log10, log2)
[1] 2.302585
> sin(pi/2) # 三角関数 (sin, cos, tan)
[1] 1
> sinpi(2/3) # sinpi(x) = sin(pi*x)
[1] 0.8660254
> acos(1/2) # 逆三角関数 (asin, acos, atan)
[1] 1.047198
```

1.2.2 数の扱い

R では実数および複素数を取り扱うことができ、指数表記にも対応している。また、無限大や不定な数など特殊なものを扱うこともできる。

```
> (1.5+3.5i) * (2-4i) # 複素数の計算
[1] 17+1i
> ## i の前に数字がある場合のみ虚数とみなすことに注意
> 1i * 1i # 虚数単位は i ではなく 1i
[1] -1+0i
> 1.38e10 * 3.68e-37 / 0.34e-5 # 指数表記の計算
[1] 1.493647e-21
> -log(0) # 無限大 (非常に大きな値)
[1] Inf
> 3 * log(0) # 数として扱える (計算はできる)
[1] -Inf
> sqrt(-1) # Not a Number (非数)
[1] NaN
> sqrt(-1) + 1 # 数として扱えないので計算はできない
[1] NaN
> log(0) / log(0) # これも Not a Number (非数)
[1] NaN
```

Rscript: `basic-numbers.r`

なお、これらの数値は C 言語にあるような int や double などの数値データの型を気にする必要はない。

1.2.3 変数への代入

文字列を変数名として、数値を保持することができる。また、変数をそのまま計算に用いることもできる。

Rscript: `basic-variables.r`

```
> x <- sin(pi/3) # x に代入
> print(x) # x の値を確認
[1] 0.8660254

> y <- cos(pi/3) # y に代入
> y # print(y) と同じ, y の値を確認
[1] 0.5

> z <- x - y # 計算結果を代入
> (z) # print(z) と同じ, z の値を確認
[1] 0.3660254

> (w <- x * y) # print(w <- x * y) と同じ, 代入結果を表示
[1] 0.4330127

> w # 代入結果を確認 (上と同じ値が表示される)
[1] 0.4330127
```

変数名は自由に決めて用いることができる(例:x, y, abcなど)。しかし、`sin`, `log`, `pi`などRの仕様として使われているものは、用いることができない訳ではないが混乱を招く元なので使わない方が良い。

なお、Rでは、変数や関数、および関数の実行結果等を総称してオブジェクト(object)と呼ぶ。

演習 1.2. R を電卓として使ってみよう。

1. 四則演算の計算順を確認する。
2. 複素数の扱い方を確認する。
3. 数学で用いられるどういった関数がRで利用可能か確認する。

1.3 データ構造

Rには、以下のようなデータ構造が用意されている。

- ベクトル (vector)
- 行列 (matrix)
- 配列 (array)
- リスト (list)
- データフレーム (data frame)

また、これらのデータは適当な変数を割り当てて保存しておくことができる。

以下ではデータ解析において基本的な役割を担うベクトル、行列、リスト、データフレームについて説明する。

1.3.1 ベクトル

ベクトルはスカラー値の集合(1次元配列)である。

スカラー値として扱われるものには、実数と複素数以外に、文字列、論理値などが含まれる。

```
> ### 実数
> (x <- 4) # 変数 x に実数 4 を代入

[1] 4

> x^10
[1] 1048576

> x^100
[1] 1.606938e+60

> x^1000 # 実数として保持できる最大値を越える

[1] Inf

> ### 複素数
> 1i # "i"の直前に数値を書く

[1] 0+1i

> (1+2i)*(2+1i)
[1] 0+5i

> try(i) # (tryを外して確認せよ)
> ## iだけでは複素数とみなされずエラーになる
> ### 文字列
> (y <- "foo") # 文字列は ' または " で括る

[1] "foo"

> (z <- "bar")
[1] "bar"

> ## "foo" や "bar" は意味のない文字列として良く用いられる
> paste(y,z) # 文字列の足し算,
[1] "foo bar"

> paste(y,z,sep="") # 区切り文字を ""(無) に指定
[1] "foobar"

> ## sep の省略時は区切り文字(separator)は " "(空白)
> try(y+z) # (確認せよ) 足し算はできずエラーになる
> ### 論理値
> TRUE # 論理値(真)

[1] TRUE

> T # 論理値(真)の省略形

[1] TRUE

> FALSE # 論理値(偽)

[1] FALSE
```

Rscript: `basic-scalar.r`

```
> F # 論理値(偽)の省略形
[1] FALSE
> as.numeric(TRUE) # as.numeric は数値に変換する関数
[1] 1
> as.numeric(F)
[1] 0
```

一般にベクトルは関数 `c()` を用いて生成することができる。ベクトルの要素を取り出すには `[]` を付けて要素の番号を指定すればよい。

これ以外に規則的な系列を生成するための関数として、等間隔の系列を作るために関数 `seq()`、繰り返しの系列を作るために関数 `rep()` などがある。また、ベクトルの長さを求めるために関数 `length()` が用意されている。

Rscript: `basic-vector.r`

```
> ### 関数 c の使い方
> (x <- c(0,1,2,3,4)) # スカラーを並べてベクトルを作成
[1] 0 1 2 3 4
> (y <- c("foo", "bar")) # 文字列のベクトル
[1] "foo" "bar"
> x[2] # ベクトルの 2 番目の要素
[1] 1
> y[2]
[1] "bar"
> x[c(1,3,5)] # 複数の要素は要素番号のベクトルで指定
[1] 0 2 4
> ### 関数 seq の使い方
> (x <- seq(0,3,by=0.5)) # 0 から 3 まで 0.5 刻みの系列
[1] 0.0 0.5 1.0 1.5 2.0 2.5 3.0
> (y <- seq(0,3,length=5)) # 0 から 3 まで長さ 5 の系列
[1] 0.00 0.75 1.50 2.25 3.00
> (z <- 1:10) # seq(1,10,by=1) と同様
[1] 1 2 3 4 5 6 7 8 9 10
> (z <- 10:1) # 10 から 1 まで 1 刻みの逆順
[1] 10 9 8 7 6 5 4 3 2 1
> z[3:8] # z の 3 番目から 8 番目の要素
[1] 8 7 6 5 4 3
> ### 関数 rep の使い方
> (x <- rep(1,7)) # 1 を 7 回繰り返す
```

```
[1] 1 1 1 1 1 1 1 1
> (y <- rep(c(1,2,3),times=3)) # (1,2,3) を 3回繰り返す
[1] 1 2 3 1 2 3 1 2 3
> (z <- rep(c(1,2,3),each=3)) # (1,2,3) を各 3回繰り返す
[1] 1 1 1 2 2 2 3 3 3
> ### その他の操作
> (x <- seq(0,2,by=0.3))
[1] 0.0 0.3 0.6 0.9 1.2 1.5 1.8
> length(x) # ベクトルの長さ
[1] 7
> y <- 2:5
> (z <- c(x,y)) # ベクトルの連結
[1] 0.0 0.3 0.6 0.9 1.2 1.5 1.8 2.0 3.0 4.0 5.0
> rev(z) # rev はベクトルを反転する関数
[1] 5.0 4.0 3.0 2.0 1.8 1.5 1.2 0.9 0.6 0.3 0.0
> LETTERS # アルファベットの大文字を要素とするベクトル
[1] "A" "B" "C" "D" "E" "F" "G" "H" "I" "J" "K" "L" "M"
[14] "N" "O" "P" "Q" "R" "S" "T" "U" "V" "W" "X" "Y" "Z"
> letters[1:10] # 小文字を要素とするベクトル
[1] "a" "b" "c" "d" "e" "f" "g" "h" "i" "j"
```

1.3.2 行列

一般に行列は関数関数 `matrix()` を用いて生成することができる。
行列の (i, j) 成分を取り出すには、`[i,j]` をつければよい。

```
> ### 関数 matrix の使い方
> x <- c(2,3,5,7,11,13) # ベクトルとして定義する
> matrix(x,2,3) # (2,3) 行列に変換する
[,1] [,2] [,3]
[1,]    2     5    11
[2,]    3     7    13

> (X <- matrix(x,ncol=3)) # 列数指定 (行数は自動的に決まる)
[,1] [,2] [,3]
[1,]    2     5    11
[2,]    3     7    13

> (Y <- matrix(x,ncol=3,byrow=TRUE)) # 横に並べる
[,1] [,2] [,3]
[1,]    2     3     5
[2,]    7    11    13

> ### その他の操作
> nrow(X) # 行数を取得する
```

Rscript: `basic-matrix.r`

```
[1] 2
> ncol(X) # 列数を取得する
[1] 3
> X[1,2] # (1,2) 成分を取り出す
[1] 5
> X[2, ] # 2行目を取り出す (列は指定しない)
[1] 3 7 13
> X[,3] # 3列目を取り出す (行は指定しない)
[1] 11 13
> as.vector(X) # ベクトル x に戻る
[1] 2 3 5 7 11 13
> as.vector(Y) # 横に並べた場合はベクトル x に戻らない
[1] 2 7 3 11 5 13
> dim(x) <- c(2,3) # ベクトルに次元属性を与えて行列化する
> x # Xと同じ型の行列になる
[,1] [,2] [,3]
[1,]    2      5     11
[2,]    3      7     13
```

1.3.3 リスト

リストは異なる構造のデータをまとめて1つのオブジェクトとして扱えるようにしたものである。リストの各要素は種類がバラバラであってもよい（例えばベクトルと行列が混在していてもよい）。一般にリストは関数 `list()` を用いて作成する。要素を取り出すには `[[[]]]` をつけて要素の番号を指定する。もしくは、各成分に名前をつけることができるので、それを用いて各成分を参照することもできる。

Rscript: `basic-list.r`

```
> ### 関数 list の使い方
> (L1 <- list(c(1,2,5,4),           # ベクトル
+               matrix(1:4,2),         # 行列
+               c("Hello","World"))) # 文字列のベクトル

[[1]]
[1] 1 2 5 4

[[2]]
[,1] [,2]
[1,]    1    3
[2,]    2    4

[[3]]
[1] "Hello" "World"

> ## 各要素のデータ型はバラバラでよい
> L1[[1]] # リスト L1 の第 1 要素を取り出す
```

```
[1] 1 2 5 4
> L1[[2]][2,1] # リストの第2要素の(2,1)成分を取り出す
[1] 2
> L1[[c(3,2)]] # リストの第3要素の2番目
[1] "World"
> L1[[3]][[2]] # 上と同じ
[1] "World"
> L1[1] # 第1要素をリストとして取り出す
[[1]]
[1] 1 2 5 4
> L1[c(1,3)] # リストの複数要素を同時に取り出す
[[1]]
[1] 1 2 5 4

[[2]]
[1] "Hello" "World"

> (L2 <- list(Info="統計データ解析",
+                 List=L1)) # 名前付きリストを生成する

$info
[1] "統計データ解析"

$list
$list[[1]]
[1] 1 2 5 4

$list[[2]]
[,1] [,2]
[1,]    1    3
[2,]    2    4

$list[[3]]
[1] "Hello" "World"

> L2[["Info"]] # 要素名で取り出す
[1] "統計データ解析"

> L2$info # 要素名で取り出す(別記法)
[1] "統計データ解析"

> names(L1) <- c("vector",
+                  "matrix",
+                  "character") # L1の要素に名前を付ける
> L1 # 変更したリストを表示する

$vector
[1] 1 2 5 4

$matrix
[,1] [,2]
[1,]    1    3
[2,]    2    4

$character
[1] "Hello" "World"
```

1.3.4 データフレーム

データフレームは同じ長さのベクトルを束ねたものであり、解析するデータを纏めた表を考えることができる。一般にデータフレームは関数 `data.frame()` を用いて作成する。要素を取り出すには `[,]` を付けて要素の行番号・列番号を指定すればよい。また、各行・各列には名前を付けることができるので、それを用いてデータを参照することもできる。

Rscript: `basic-data.frame.r`

```
> ### 関数 data.frame の使い方
> (x <- data.frame( # 各項目が同じ長さのベクトルを並べる
+   month=c(4,5,6,7),           # 月
+   price=c(900,1000,1200,1100), # 價格
+   deal=c(100,80,50,75)))     # 取引量

  month price deal
1     4    900   100
2     5   1000    80
3     6   1200    50
4     7   1100    75

> x[2,3]  # 2行3列を取り出す
[1] 80

> x[3, ]  # 3行目を取り出す
  month price deal
3     6   1200    50

> x[,2]  # 2列目を取り出す
[1] 900 1000 1200 1100

> x$price # 列名で取り出す (上記の別記法)
[1] 900 1000 1200 1100

> x[2]      # 2列目だけからなるデータフレームを取り出す
  price
1    900
2   1000
3   1200
4   1100

> x["price"] # 列名で取り出す (上記の別記法)
  price
1    900
2   1000
3   1200
4   1100

> x[c("month", "deal")] # 複数列の場合はベクトルで指定する
  month deal
1     4   100
2     5    80
3     6    50
4     7    75

> ### 行・列の名前の操作
> rownames(x) # 行の名前を表示する
```

```
[1] "1" "2" "3" "4"

> rownames(x) <- c("Apr", "May", "Jun", "Jul") # 上書き
> colnames(x) # 列の名前を表示する

[1] "month" "price" "deal"

> colnames(x) <- c("tsuki", "kakaku", "torihiki") # 上書き
> x # 変更されたデータフレームを表示する

  tsuki kakaku torihiki
Apr      4     900     100
May      5    1000      80
Jun      6    1200      50
Jul      7    1100      75

> x["May", "kakaku"] # 特定の要素を名前で参照する

[1] 1000
```

演習 1.3. 実際のデータに基づいてデータフレームを作成してみよう。

1. 長さの等しいベクトルを作成する。
2. ベクトルを束ねてデータフレームを作成する。
3. データフレームの行・列に適当な名前に変更する。

1.4 補遺

1.4.1 参考文献

確率論、統計学および R の操作に関する成書は多数あるが、以下を参考として挙げておく。これ以外にも多数あるので、図書館などで手に取って自分に合ったものを選ばれたい。

- [1] 藤澤洋徳. **確率と統計**. 東京: 朝倉書店, 2006.
- [2] 吉田朋広. **数理統計学**. 東京: 朝倉書店, 2006.
- [3] 竹内啓. **数理統計学**. 東京: 東洋経済, 1963.
- [4] 金明哲. **Rによるデータサイエンス(第2版)**. 東京: 森北出版, 2017.
- [5] U. リゲス (石田基広訳). **Rの基礎とプログラミング技法**. 東京: 丸善出版, 2012.
- [6] 奥村晴彦. **Rで楽しむ統計**. 東京: 共立出版, 2016.
- [7] Larry Wasserman. *All of Statistics*. New York: Springer, 2004.
- [8] Gareth James et al. *An Introduction to Statistical Learning with Applications in R*. New York: Springer, 2013.

1.4.2 亂数の生成

代表的な確率分布に従う乱数の生成を行うことができる。ここでは一様乱数、正規乱数およびランダムサンプリングを行う関数 `runif()`, `rnorm()`, `sample()` を紹介する。

Rscript: `basic-random.r`

```
> ### 関数 runif の使い方
> runif(4,min=-1,max=1) # [-1,1] 上の一様乱数を 4 個生成
[1] -0.10998325 -0.12918400  0.08723451 -0.90943464
> runif(4) # 最大最小の指定がなければ [0,1] 上の一様分布
[1] 0.5291765 0.8421631 0.5324661 0.2800179
> ### 関数 rnorm の使い方
> rnorm(4,mean=3,sd=2) # 平均 3, 標準偏差 2 の正規乱数
[1] 0.2088095 1.9740474 3.1575775 4.1950552
> rnorm(4) # 平均と標準偏差の指定がなければ標準正規分布
[1] 1.38033958 -0.42721086 0.10326028 -0.06013144
> ### 関数 sample の使い方
> sample(1:10,size=5) # 1-10 の整数からランダムに 5 つ抽出
[1] 3 10 7 9 4
> sample(1:10,10) # ランダムに並べ替え。"size=" は省略可
[1] 2 4 5 1 6 9 8 7 10 3
> sample(1:10,10,replace=TRUE) # 復元抽出
[1] 7 1 3 4 8 2 2 7 2 8
```

これらの関数は数値シミュレーションを行う場合に重要な役割を果たす。

演習 1.4. R を使って乱数を生成してみよう。

1. 計算機で生成される乱数の性質を確認しよう。
(ヒント: `help("Random")`)
2. 乱数を複数生成し、そのちらばり具合を確認してみよう。
(ヒント: `help("summary")`, `help("hist")`)

ベクトル・行列の演算と関数

データ解析、多変量解析、パターン認識などで必要となる計算の多くはベクトルと行列を用いた計算である。この章では、これらを **R** 言語で実現する方法をまとめます。

2.1 ベクトルの計算

まずベクトルのみでのさまざまな計算をまとめます。以下ではベクトルを太字で、その要素は下付き添字で表現する。例えば k 次元ベクトルは

$$\mathbf{a} = (a_1, a_2, \dots, a_k)$$

のように表す。またベクトル \mathbf{a} の第 i 成分を指す場合には $(\mathbf{a})_i$ のように書くこともある。

2.1.1 和

同じ長さのベクトルの和および差

$$\mathbf{a} \pm \mathbf{b} = (a_1 \pm b_1, a_2 \pm b_2, \dots, a_k \pm b_k)$$

は、数値の和と差のように扱うことができる。成分による表現では

$$(\mathbf{a} \pm \mathbf{b})_i = a_i \pm b_i$$

と書くことができる。

```
> # ベクトルの加減・スカラー倍
> a <- 1:3 # 長さ 3 のベクトル
> b <- 4:6 # 長さ 3 のベクトル
> a + b # 足し算

[1] 5 7 9

> a - b # 引き算も可
[1] -3 -3 -3

> 2 * a # スカラー倍
[1] 2 4 6

> 2*a - b/3 # 線形結合
[1] 0.6666667 2.3333333 4.0000000

> a + 1:6 # 長さが異なる場合は短い方を反復 c(1:3,1:3)+1:6
[1] 2 4 6 5 7 9

> a + 1:5 # 長さが整数倍でない場合は警告される
[1] 2 4 6 5 7
```

Rscript: `vector-sum.r`

2.1.2 積

ベクトルの積は通常内積

$$\mathbf{a} \cdot \mathbf{b} = \sum_{i=1}^k a_i b_i$$

を指すが、データ解析においては要素毎の積 (Hadamard product, Schur product)

$$\mathbf{a} \circ \mathbf{b} = (a_1 b_1, a_2 b_2, \dots, a_k b_k)$$

すなわち

$$(\mathbf{a} \circ \mathbf{b})_i = a_i b_i$$

を計算する場合も多い。2つの意味での積が簡単に計算できるよう二項演算子 `%*%` (内積) および `*` (要素毎の積) が定義されている。

Rscript: `vector-prod.r`

```
> a <- 1:3 # 長さ 3 のベクトル
> b <- 4:6
> a %*% b # ベクトルの内積 (計算結果は 1x1 行列)

[,1]
[1,] 32

> try(a %*% (1:6)) # (確認せよ) 長さが異なるとエラーとなる
> a * b # 要素毎の積 (計算結果はベクトル)

[1] 4 10 18

> a * 1:6 # 長さが異なる場合は足りない方が周期的に拡張される

[1] 1 4 9 4 10 18

> a / b # 除算も成分ごとに計算される

[1] 0.25 0.40 0.50
```

2.1.3 初等関数の適用

ベクトルに初等関数 (`sin`, `exp`, … など) を適用すると、成分ごとに計算した結果が返される。例えば、ベクトル \mathbf{a} に関数 `sin` を適用した結果は

$$\sin(\mathbf{a}) = (\sin(a_1), \dots, \sin(a_k))$$

となる。

Rscript: `vector-fun.r`

```
> a <- (1:6) * pi/2 # 長さ 6 のベクトル
> sin(a) # 数値誤差のため正確に 0 とならない成分がある

[1] 1.000000e+00 1.224647e-16 -1.000000e+00
[4] -2.449294e-16 1.000000e+00 3.673940e-16

> exp(a)
```

```
[1] 4.810477 23.140693 111.317778 535.491656
[5] 2575.970497 12391.647808

> log(a)

[1] 0.4515827 1.1447299 1.5501950 1.8378771 2.0610206
[6] 2.2433422
```

演習 2.1. ベクトルの計算をしてみよう.

1. ベクトルの内積から 2 つのベクトルがなす角を求めよ.
2. 2 次元および 3 次元ベクトルの積としては、これら以外に“外積”がある。どのように計算すればよいか調べよ。

2.2 行列とその演算

次に行列のみでのさまざまな計算をまとめる。以下では行列を大文字で、その要素は下付き添字で表現する。例えば $m \times n$ 行列は

$$A = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & & \ddots & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{pmatrix}$$

のように表す。また、行列 A の (i, j) 成分を指す場合には $(A)_{ij}$ のように書くこともある。

2.2.1 和

同じ大きさの行列の和および差

$$(A \pm B)_{ij} = a_{ij} \pm b_{ij}$$

は、ベクトルと同じように記述することができる。

```
> (A <- matrix(1:6,nrow=2,ncol=3)) # 2x3 行列の作成
      [,1] [,2] [,3]
[1,]    1    3    5
[2,]    2    4    6

> (B <- rbind(c(2,3,5),c(7,11,13))) # row bind
      [,1] [,2] [,3]
[1,]    2    3    5
[2,]    7   11   13

> ## 行ベクトル (row) として連結
> (C <- cbind(c(0,0),c(0,1),c(1,0))) # column bind
      [,1] [,2] [,3]
[1,]    0    0    1
[2,]    0    1    0

> ## 列ベクトル (column) として連結
> A + B - C
```

Rscript: `matrix-sum.r`

```
[,1] [,2] [,3]
[1,]    3    6    9
[2,]    9   14   19
```

2.2.2 積

行列の積

$$(AB)_{ij} = \sum_{k=1}^m a_{ik} b_{kj}$$

は、左側の行列の行ベクトルと右側の行列の列ベクトルの内積を各要素とする行列となるので、左側の行列の行数と右側の行列の列数が一致する場合のみ定義される。この積は二項演算子 `%*%` を用いて計算する。なお、行列の転置 (transpose) は関数 `t()` を用いて計算することができ、ある行列とその転置行列の積が簡単に計算できる。これは分散などの計算に活躍する。

一方、ベクトルと同様に同じ大きさの行列の要素毎の積 (Hadamard product, Schur product)

$$(A \circ B)_{ij} = a_{ij} b_{ij}$$

も簡単に計算できるようになっており、これは二項演算子 `*` を用いて計算する。

Rscript: `matrix-prod.r`

```
> (A <- matrix(1:6,nrow=2,ncol=3)) # 行列の作成 (2x3行列)
[,1] [,2] [,3]
[1,]    1    3    5
[2,]    2    4    6

> (B <- rbind(c(2,3,5),c(7,11,13))) # 行ベクトルを連結
[,1] [,2] [,3]
[1,]    2    3    5
[2,]    7   11   13

> (C <- cbind(c(2,3,5),c(7,11,13))) # 列ベクトルを連結
[,1] [,2]
[1,]    2    7
[2,]    3   11
[3,]    5   13

> A * B # 要素ごとの積
[,1] [,2] [,3]
[1,]    2    9   25
[2,]   14   44   78

> A / B # 除算も成分ごと
[,1]      [,2]      [,3]
[1,] 0.5000000 1.0000000 1.0000000
[2,] 0.2857143 0.3636364 0.4615385

> A %*% C # 行列の積 (結果は 2x2 行列)
```

```
[,1] [,2]
[1,] 36 105
[2,] 46 136

> C %*% B # 行列の積 (結果は 3x3 行列)

[,1] [,2] [,3]
[1,] 53 83 101
[2,] 83 130 158
[3,] 101 158 194

> A %*% t(A) # 行列 A とその転置行列の積 (結果は 2x2 行列)

[,1] [,2]
[1,] 35 44
[2,] 44 56
```

2.2.3 初等関数の適用

ベクトルの場合と同様に、行列に初等関数 (\sin , \exp , … など) を適用すると、成分ごとに計算した結果が返される。例えば、行列 A に関数 \sin を適用した結果は

$$\sin(A) = \begin{pmatrix} \sin(a_{11}) & \sin(a_{12}) & \dots & \sin(a_{1n}) \\ \sin(a_{21}) & \sin(a_{22}) & \dots & \sin(a_{2n}) \\ \vdots & & \ddots & \vdots \\ \sin(a_{m1}) & \sin(a_{m2}) & \dots & \sin(a_{mn}) \end{pmatrix}$$

で与えられ、行列 A と同じサイズの行列となる。

```
> A <- matrix((1:6)*pi/2, 2, 3) # 行列の作成 (2x3 行列)
> sin(A)

[,1]          [,2]          [,3]
[1,] 1.000000e+00 -1.000000e+00 1.000000e+00
[2,] 1.224647e-16 -2.449294e-16 3.67394e-16

> exp(A)

[,1]          [,2]          [,3]
[1,] 4.810477 111.3178 2575.97
[2,] 23.140693 535.4917 12391.65

> log(A)

[,1]          [,2]          [,3]
[1,] 0.4515827 1.550195 2.061021
[2,] 1.1447299 1.837877 2.243342
```

Rscript: `matrix-fun.r`

2.2.4 行列式とトレース

行列に特有な量として行列式とトレース(対角成分の総和)があるが、行列式は関数 `det()` を用いて計算することができる。一方、トレースは専用の関数は用意されていないが、対角成分を取り出す関数 `diag()` とベクトルの和を計算する関数 `sum()` を用いて簡単に計算できる。

Rscript: matrix-det.r

```
> (A <- matrix(1:9, nrow=3, ncol=3)) # 行列の作成 (3x3 行列)
      [,1] [,2] [,3]
[1,]    1    4    7
[2,]    2    5    8
[3,]    3    6    9

> det(A) # 行列式 (determinant) の計算
[1] 0

> sum(diag(A)) # トレス (trace) の計算
[1] 15
```

2.2.5 逆行列

正方行列の逆行列 (inverse matrix) を求めるには関数 `solve()` を用いる。

Rscript: matrix-inv.r

```
> (A <- matrix(c(2,3,5,7,11,13,17,19,23),
+                           nrow=3, ncol=3)) # 正則な正方行列 (3x3)
      [,1] [,2] [,3]
[1,]    2    7   17
[2,]    3   11   19
[3,]    5   13   23

> (B <- solve(A)) # 逆行列の計算
      [,1]      [,2]      [,3]
[1,] -0.07692308 -0.7692308  0.69230769
[2,] -0.33333333  0.5000000 -0.16666667
[3,]  0.20512821 -0.1153846 -0.01282051

> A %*% B # AB = BA = E(単位行列) となることを確認する
      [,1]      [,2]      [,3]
[1,]  1.000000e+00   0 -1.110223e-16
[2,] -4.440892e-16   1 -1.942890e-16
[3,]  0.000000e+00   0  1.000000e+00

> B %*% A
      [,1]      [,2]      [,3]
[1,]  1.000000e+00 3.552714e-15 1.776357e-15
[2,] -2.220446e-16 1.000000e+00 0.000000e+00
[3,]  0.000000e+00 -1.387779e-16 1.000000e+00
```

演習 2.2. 行列の計算をしてみよう。

1. 単位行列と成分が全て 1 の行列を作成せよ。
2. 適当な 2 次正方行列 A に対して, Hamilton-Cayley の定理

$$A^2 - \text{tr}(A)A + \det(A)E_2 = O$$

の成立を確認せよ。ただし E_2 は 2 次単位行列, O は 2 次正方零行列であり, また $\text{tr}(A), \det(A)$ はそれぞれ A のトレス, 行列式を表す。

2.3 ベクトルと行列の計算

2.3.1 行列とベクトルの積

R言語においては、列ベクトル・行ベクトルという区別はなく、どちらも同じベクトルとして扱われる。行列とベクトルの積においては、行列のどちらからベクトルを掛けるかによって自動的に列ベクトルか行ベクトルが判断されて扱われる。なお、ベクトルも行列の一種であるから、計算結果は行列として表現されることに注意する。

```
> (A <- matrix(1:16, nrow=4, ncol=4))

[,1] [,2] [,3] [,4]
[1,]    1     5     9    13
[2,]    2     6    10    14
[3,]    3     7    11    15
[4,]    4     8    12    16

> (b <- rnorm(4))

[1] -0.01592495  1.56170064  0.23708219  0.79810672

> A %*% b # 列ベクトルとして計算

[,1]
[1,] 20.30171
[2,] 22.88267
[3,] 25.46363
[4,] 28.04460

> b %*% A # 行ベクトルとして計算

[,1]      [,2]      [,3]      [,4]
[1,] 7.01115 17.33501 27.65887 37.98272
```

Rscript: linear-calc.r

2.3.2 連立方程式

データ解析の様々な場面で連立一次方程式が現れるが、これは行列とベクトルで表現される。逆行列の計算に用いた関数 `solve()` の引数として行列とベクトルを与えることによって連立一次方程式を解くことができる。

```
> ### 連立一次方程式
> (A <- matrix(c(2,3,5,7,11,13,17,19,23), 3, 3)) # 3x3 行列

[,1] [,2] [,3]
[1,]    2     7    17
[2,]    3    11    19
[3,]    5    13    23

> (b <- c(2,3,5))

[1] 2 3 5

> (x <- solve(A, b)) # Ax = b を解く

[1] 1.000000e+00 -2.220446e-16  5.124106e-17

> A %*% x # Ax が b と一致するか確認する
```

Rscript: linear-eq.r

```

[,1]
[1,]    2
[2,]    3
[3,]    5

> ### 行列方程式 (A が同じ複数の連立一次方程式と考えられる)
> (B <- matrix(c(2,1,1,1),2,2)) # 2x2 行列

[,1] [,2]
[1,]    2    1
[2,]    1    1

> (C <- matrix(c(4,3,10,7),2,2)) # 2x2 行列

[,1] [,2]
[1,]    4   10
[2,]    3    7

> (X <- solve(B,C)) # BX = C を解く

[,1] [,2]
[1,]    1    3
[2,]    2    4

> B%*%X # BX が C と一致するか確認する

[,1] [,2]
[1,]    4   10
[2,]    3    7

```

演習 2.3. 行列とベクトルの計算をしてみよう.

1. 適当な 2 次元のベクトルを 120 度回転しなさい.
2. $n \times n$ 行列 A と n 次のベクトル b を作成し, $A \% * \% b + b \% * \% A$ を計算せよ (エラーになる). 何故, そうなるか理由を考えなさい.
3. 連立方程式の問題を作成し, それを解きなさい.

2.4 関数

2.4.1 制御文

一般に最適化や数値計算などを行うためには, 条件分岐や繰り返しを行うための仕組みが必要となる. 多くの計算機言語では `if`(条件分岐), `for`・`while`(繰り返し)を用いた構文が用意されているが, これを制御文と言う. R 言語においてもこれらの構文は用意されており, 制御文を使うことによって次節で述べるような複雑な計算を行う関数を定義することができる.

Rscript: `fun-control.r`

```

> ### 条件分岐 (if)
> x <- 5
> if(x > 0) { # 正か否か判定
+   ## 条件が真の場合に実行するブロック
+   print("positive")
+ } else {
+   ## 条件が偽の場合に実行するブロック
+   print("negative")

```

```

+ }

[1] "positive"

> ## else 以下はなくとも動く
> if(x > 0) {
+   print("positive")
+ }

[1] "positive"

> ## 評価が簡便な場合の条件分岐 (ifelse)
> ifelse(x < 0, "true", "not true")

[1] "not true"

> ### 繰り返し (for)
> y <- 0
> for(i in 1:10) { # 1-10 の合計を計算
+   y <- y + i
+ }
> print(y)

[1] 55

> ### 繰り返し (while)
> z <- 1
> n <- 0
> while(z < 100) { # 100未満の間は 2倍し続ける
+   z <- 2 * z
+   n <- n + 1
+ }
> print(z) # 100を超えた際の z の値

[1] 128

> print(n) # 条件を満たすまでの回数

[1] 7

```

2.4.2 関数の定義

一般に関数とは入力を規則に従って変換し出力する仕組みを指す。Rでは、入力を引数(argument)、出力を返値(value)と呼び、関数function()を用いて自由に関数を定義することができる。

```

> ### 階乗を計算する関数
> fact <- function(n){ # 素直に計算
+   ifelse(n>0,prod(1:n),1)
+ }

> fact2 <- function(n){ # 再帰的に定義
+   if(n>0) {
+     return(n*fact2(n-1)) # 自分を呼び出す
+   } else {
+     return(1) # fact2(0) = 0! = 1
+   }
+ }

> fact(10) # 同じ結果になるか確認する

[1] 3628800

> fact2(10)

```

Rscript: `fun-define.r`

```
[1] 3628800
```

同じ機能を持つ関数でも定義の仕方はいろいろ工夫できる。

演習 2.4. 以下の機能を持つ関数を作ってみよう。

1. 2次方程式の3つの係数を入力すると解を出力する関数を作成しなさい。
2. 第1項, 第2項, および項数を入力すると Fibonacci 数列を指定された項数まで出力する関数を作成しなさい。なお, Fibonacci 数列とは下記の漸化式を満たす数列である。

$$a_n = a_{n-1} + a_{n-2}, \quad n = 3, 4, \dots$$

3. 自身で仕様を決めて, それを満たす関数を作成しなさい。

2.5 補遺

2.5.1 参考文献

この章の内容に関連する参考書として以下を挙げておく。

- [1] U. リゲス (石田基広訳). *R の基礎とプログラミング技法*. 東京: 丸善出版, 2012.

2.5.2 ベクトルのノルム

通常の l^2 ノルム

$$\|\mathbf{a}\|_2 = \sqrt{\sum_{i=1}^k |a_i|^2}$$

は内積を用いれば計算できる。一般の l^p ノルム

$$\|\mathbf{a}\|_p = \left(\sum_{i=1}^k |a_i|^p \right)^{1/p}$$

や l^∞ ノルム

$$\|\mathbf{a}\|_\infty = \max_{1 \leq i \leq k} |a_i|$$

は, 要素の和を計算する関数 `sum()`, および最大値を取り出す関数 `max()` を利用して計算することができる。

Rscript: `vector-norm.r`

```
> (a <- rnorm(6)) # 標準正規乱数でベクトルを作成
[1] -0.9758500  0.6328807  1.4675893  0.7444935
[5] -0.1985488 -0.2357815
> sqrt(a %*% a) # l2 ノルム (計算結果は行列型で表示される)
```

```
[,1]
[1,] 2.038608

> as.vector(sqrt(a %*% a)) # l2 ノルム (ベクトル型に変換)
[1] 2.038608

> sum(abs(a)) # l1 ノルム
[1] 4.255144

> sum(abs(a)^2)^(1/2) # lp ノルム (p=2)
[1] 2.038608

> sum(abs(a)^3)^(1/3) # lp ノルム (p=3)
[1] 1.6842

> max(abs(a)) # 最大ノルム
[1] 1.467589
```

2.5.3 行列のノルム

行列のノルムにはいくつか定義があるが、関数 `norm()` によって作用素ノルム ($p = 1$ および $p = \infty$),

$$\|A\|_1 = \max_{1 \leq j \leq n} \sum_{i=1}^m |a_{ij}|$$

$$\|A\|_\infty = \max_{1 \leq i \leq m} \sum_{j=1}^n |a_{ij}|$$

Frobenius ノルム,

$$\|A\|_F = \sqrt{\sum_{i=1}^m \sum_{j=1}^n |a_{ij}|^2}$$

最大ノルム,

$$\|A\|_{\max} = \max\{|a_{ij}|\}$$

およびスペクトルノルム

$$\|A\|_2 = \sigma_{\max}(A), \quad (A \text{ の最大特異値})$$

を計算することができる。

```
> (A <- matrix(c(1,0,0,0,2,0,1,-1,-3,0,0,0),
+                  nrow=3,ncol=4)) # 3x4 行列

[,1] [,2] [,3] [,4]
[1,]    1    0    1    0
[2,]    0    2   -1    0
[3,]    0    0   -3    0

> norm(A,type="O") # 作用素ノルム (p=1; one norm)
```

Rscript: `matrix-norm.r`

```
[1] 5
> norm(A,type="I") # 作用素ノルム (infinity norm)
[1] 3
> norm(A,type="F") # Frobenius ノルム (Frobenius norm)
[1] 4
> norm(A,type="M") # 最大ノルム (max norm)
[1] 3
> norm(A,type="2") # スペクトルノルム (spectral/2-norm)
[1] 3.408689
```

2.5.4 一般化逆行列

データ解析においては、正方でない行列の一般化逆行列(擬似逆行列; pseudo-inverse matrix)がしばしば必要となる。一般化逆行列 A^\dagger とは

$$AA^\dagger A = A$$

が成り立つ行列のことである。いくつかの定義がある。一般化逆行列の中で良く用いられるのは Moore-Penrose の一般化逆行列(Moore-Penrose pseudoinverse)と呼ばれるもので、これを A^+ と書くことにすると

$$\begin{aligned} AA^+A &= A \\ A^+AA^+ &= A^+ \\ (AA^+)^* &= AA^+ \quad (* \text{は随伴行列の意}) \\ (A^+A)^* &= A^+A \end{aligned}$$

が成立する行列である。ただし、随伴行列とは、転置かつ複素共役をとった行列のことである。これはパッケージ MASS の中の関数 `ginv()` を用いて求めることができる。

Rscript: `matrix-ginv.r`

```
> ### 一般化逆行列
> library(MASS) # MASS パッケージを読み込む
> (C <- matrix(rnorm(6),
+               nrow=2,ncol=3)) # ランダムに 2x3 行列を作成
      [,1]      [,2]      [,3]
[1,] -0.4781938  0.1507989  0.1250757
[2,]  0.6354841 -1.1556820  0.3683875

> (D <- ginv(C)) # 一般化逆行列の計算
      [,1]      [,2]
[1,] -1.9807833 -0.1175267
[2,] -0.6895964 -0.7752154
[3,]  1.2535775  0.4853147

> C %*% D %*% C # CC^+ + C = C であることを確認
```

```
[,1]      [,2]      [,3]
[1,] -0.4781938  0.1507989  0.1250757
[2,]  0.6354841 -1.1556820  0.3683875
```

2.5.5 固有値と固有ベクトル

一般に n 次正方行列 A に対して、複素数 λ と零ベクトルでない n 次元ベクトル x が

$$Ax = \lambda x$$

を満たすとき、 λ を A の固有値、 x を λ に対する固有ベクトルと呼ぶ。固有値および固有ベクトルはデータ解析にしばしば用いられ、R では関数 `eigen()` で求めることができる。

```
> (A <- matrix(c(1,-1,-1,1),2,2)) # 2x2行列
[,1] [,2]
[1,]    1   -1
[2,]   -1    1

> r <- eigen(A) # 結果は固有値と固有ベクトルからなるリスト
> r$values # 固有値
[1] 2 0

> r$vectors # 固有ベクトルからなる行列
[,1]      [,2]
[1,] -0.7071068 -0.7071068
[2,]  0.7071068 -0.7071068

> ## r$vectors[,i] が r$values[i] に対する固有ベクトル
> t(r$vectors) %*% A %*% r$vectors # 対角化
[,1] [,2]
[1,]    2    0
[2,]    0    0
```

Rscript: `matrix-eigen.r`

2.5.6 特異値分解

大規模データに対するデータ解析では、正方行列でない任意の行列に対する分解が必要となることがある。一般に実 $n \times p$ 行列 A に対して、 $q = \min\{n, p\}$ とすると、実 $n \times q$ 行列 U 、実 $p \times q$ 行列 V 、非負 q 次対角行列 D が存在して、 $U^T U = V^T V = E_q$ を満たし、かつ

$$A = UDV^T$$

と書けることが知られている。ただし T は転置の意で、 E_q は q 次単位行列を表す。この分解を A の特異値分解と呼び、 D の対角成分を A の特異値と呼ぶ。また、行列 A の特異値は A から一意的に定まることが知られている。特異値分解は関数 `svd()` で実行できる。

Rscript: matrix-svd.r

```

> (A <- matrix(1:6,nrow=2)) # 2x3 行列
      [,1] [,2] [,3]
[1,]    1    3    5
[2,]    2    4    6

> s <- svd(A) # 結果は特異値と行列 U,V からなるリスト
> s$d # 特異値
[1] 9.5255181 0.5143006

> s$u # 行列 U
      [,1]          [,2]
[1,] -0.6196295 -0.7848945
[2,] -0.7848945  0.6196295

> s$v # 行列 V
      [,1]          [,2]
[1,] -0.2298477  0.8834610
[2,] -0.5247448  0.2407825
[3,] -0.8196419 -0.4018960

> s$u %*% diag(s$d) %*% t(s$v) # 行列 A の再現
      [,1] [,2] [,3]
[1,]    1    3    5
[2,]    2    4    6

```

演習 2.5.

以下の計算をしてみよう。

- 関数 `ginv()` で計算される行列が Moore-Penrose の一般化逆行列になっていることを確かめよ。
- 行列 A の Frobenius ノルムを, $A * A$ および $A \% * \% t(A)$ を用いて計算しなさい。 (ヒント: $\|A\|_F^2 = \text{tr } AA^T$)
- A が非負定値 n 次対称行列, すなわち固有値がすべて非負の n 次対称行列のとき, A の対角化を考えることで $A = B^2$ を満たす非負定値 n 次対称行列 B が求められる。この行列 B を計算するプログラムを作成せよ。

データの加工・整理と入出力

収集されたデータを整理して実際の解析を行うためには、特定の条件に当て嵌まる行や列の選択、複数の量から計算した統計量によって新たな行を作成、データをグループ化して集計など、様々な操作が要請される。以下では、基本的なデータの型とその操作について解説する。

3.1 データの形式

まず、Rで用いられる基本的なデータ形式についてもう一度まとめておく。

3.1.1 値の型

Rでは数値、文字列、真偽値を扱うことができる。どの型であるかを確認する場合には関数`mode()`や関数`typeof()`(数値の型などについて少し詳しい)を用いればよい。

```
> ### 値の型
> (a <- 2.718)      # 数値 (実数)
[1] 2.718
> mode(a)           # 型を確認
[1] "numeric"
> typeof(a)          # 型を確認 (内部での詳しい分類)
[1] "double"
> (b <- 3.5 + 5.8i) # 数値 (複素数)
[1] 3.5+5.8i
> mode(b)           # 型を確認
[1] "complex"
> (c <- "alphabet") # 文字列
[1] "alphabet"
> mode(c)           # 型を確認
[1] "character"
> (d <- FALSE)       # 真偽値 (TRUE/FALSE)
[1] FALSE
> mode(d)           # 型を確認
[1] "logical"
```

Rscript: [data-type.r](#)

3.1.2 ベクトル

1つ、または複数の値を並べたものがベクトルである。規則的な系列を生成するためには関数 `seq()` や関数 `rep()` など、いろいろな方法が用意されている。ベクトルの長さを知るためにには関数 `length()` を用いる。

Rscript: `data-vector.r`

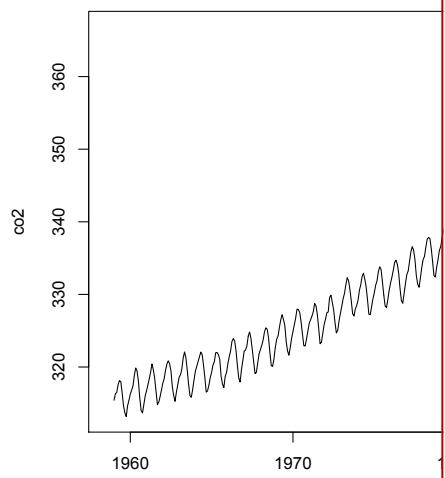


図 3.1: ベクトルデータの表示の例

図 3.1 参照

```
> ### ベクトルの例
> (a <- c(2,3,5,7,11)) # 要素を指定
[1] 2 3 5 7 11
> (b <- 7.5:1) # 差 1 の等差数列 (seq(7.5,1,by=-1)と同じ)
[1] 7.5 6.5 5.5 4.5 3.5 2.5 1.5
> (c <- rep(c(2,3),length.out=7)) # 長さ 7まで繰り返し
[1] 2 3 2 3 2 3 2
> ### データ例 datasets::co2 (詳細は help(co2) を参照)
> length(co2) # 長さを確認
[1] 468
> head(co2,n=8) # 最初の 8 個を表示
[1] 315.42 316.31 316.50 317.56 318.13 318.00 316.39
[8] 314.65
> plot(co2) # グラフ表示 (ベクトル+時系列の情報)
```

3.1.3 行列

行列は 2 次元状に値を並べたものであり、次節で説明する配列の 2 次元版と考えることができる。行列は、関数 `matrix()` によって 1 つのベクトルを並べ替える、あるいは関数 `cbind()` や関数 `rbind()` によって複数のベクトルを連結して作成することができる。行列の次元(サイズ)を知るためにには関数 `dim()` を用い、行数および列数を求めるにはそれぞれ関数 `nrow()`、関数 `ncol()` を用いる。また、関数 `rownames()` および関数 `colnames()` を用いて行と列に名前を付けることができる。

Rscript: `data-matrix.r`

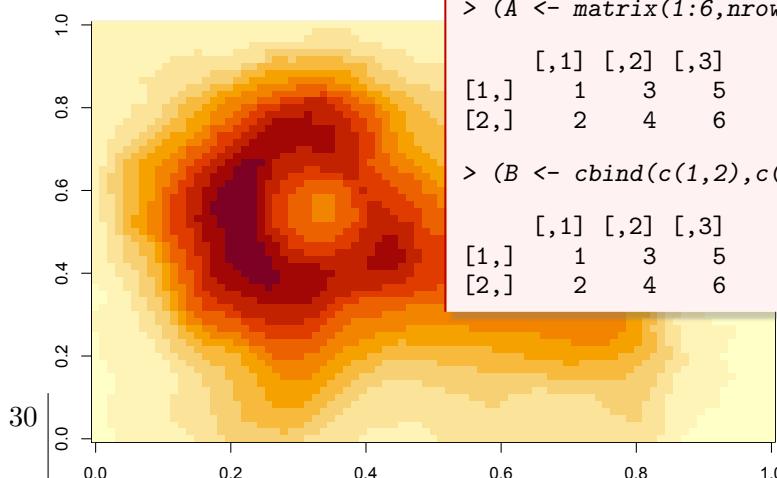


図 3.2 参照

```
> ### 行列の例 (ベクトルから行列を作成)
> (A <- matrix(1:6,nrow=2,ncol=3)) # 列ごとに並べる
[,1] [,2] [,3]
[1,]    1    3    5
[2,]    2    4    6
> (B <- cbind(c(1,2),c(3,4),c(5,6))) # 列で結合する
[,1] [,2] [,3]
[1,]    1    3    5
[2,]    2    4    6
```

```

> (C <- matrix(1:6,nrow=2,ncol=3,
+               byrow=TRUE)) # 行ごとに並べる
      [,1] [,2] [,3]
[1,]    1    2    3
[2,]    4    5    6

> (D <- rbind(c(1,2,3),c(4,5,6))) # 行で結合する
      [,1] [,2] [,3]
[1,]    1    2    3
[2,]    4    5    6

> dim(D) # 大きさを確認
[1] 2 3

> (rownames(D) <- LETTERS[1:nrow(D)]) # 行に名前を付ける
[1] "A" "B"

> (colnames(D) <- letters[1:ncol(D)]) # 列に名前を付ける
[1] "a" "b" "c"

> D # D の内容を確認する
     a b c
A 1 2 3
B 4 5 6

> ### データ例 datasets::volcano (詳細は help(volcano))
> dim(volcano) # 大きさを確認
[1] 87 61

> volcano[1:3,1:5] # 左上の 3 行 5 列を表示
      [,1] [,2] [,3] [,4] [,5]
[1,]   100   100   101   101   101
[2,]   101   101   102   102   102
[3,]   102   102   103   103   103

> image(volcano) # 濃淡図として表示

```

3.1.4 配列

行列を一般化したものとして配列が用意されており、関数 `array()` を用いてベクトルを並べ替え作成することができる。次元を知るためにには行列と同様に関数 `dim()` を用いる。

図 3.3 参照

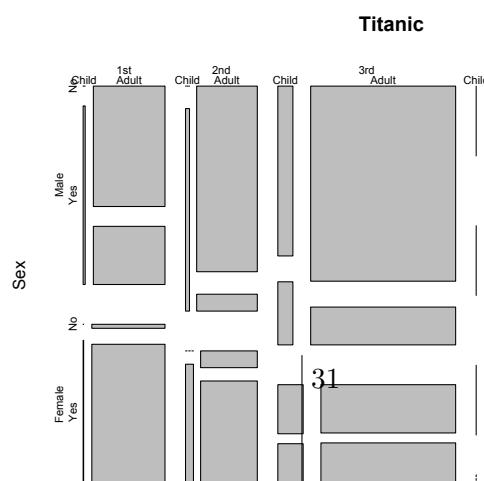
```

> ### 配列の例
> (A <- array(1:13,dim=c(3,4,2))) # 3x4x2 次の配列
, , 1

      [,1] [,2] [,3] [,4]
[1,]    1    4    7   10
[2,]    2    5    8   11
[3,]    3    6    9   12

```

Rscript: [data-array.r](#)



```

, , 2

[,1] [,2] [,3] [,4]
[1,]   13    3    6    9
[2,]    1    4    7   10
[3,]    2    5    8   11

> ### データ例 datasets::Titanic (詳細は help(Titanic))
> dim(Titanic) # 大きさを確認

[1] 4 2 2

> Titanic      # データを表示

, , Age = Child, Survived = No

Sex
Class Male Female
1st    0     0
2nd    0     0
3rd   35    17
Crew   0     0

, , Age = Adult, Survived = No

Sex
Class Male Female
1st  118     4
2nd  154    13
3rd  387    89
Crew 670     3

, , Age = Child, Survived = Yes

Sex
Class Male Female
1st    5     1
2nd   11    13
3rd   13    14
Crew   0     0

, , Age = Adult, Survived = Yes

Sex
Class Male Female
1st   57   140
2nd   14    80
3rd   75    76
Crew  192   20

> plot(Titanic) # タイプ図として表示

```

3.1.5 データフレーム

データフレームは表を取り扱うためのデータ形式で、行列と同様な2次元の配列と考えることができる。関数 `data.frame()` によってベクトルを連結して作成する、あるいは行列を変換して作成するなど、さまざまな形式のデータから変換して作成する方法が用意されている。また、データフレームの中から必要な部分集合を

取り出すために、関数 `subset()` が用意されている。行列と同様に行名、列名をつけることができるが、行名には関数 `rownames()` を、列名には関数 `names()` を用いる。

図 3.4 参照

```
> ### データフレームの例 (ベクトルから行列を作成)
> (A <- data.frame(height=c(172, 158, 160),
+                     weight=c(60, 53, 51)))

  height weight
1    172     60
2    158     53
3    160     51

> (B <- matrix(1:8, nrow=4, ncol=2))

 [,1] [,2]
[1,]    1    5
[2,]    2    6
[3,]    3    7
[4,]    4    8

> (C <- data.frame(B)) # 行列から作ることもできる

  X1 X2
1  1  5
2  2  6
3  3  7
4  4  8

> (rownames(C) <- letters[1:nrow(C)]) # 行名を付ける
[1] "a" "b" "c" "d"

> (names(C) <- c("Left", "Right")) # 列名を付ける
[1] "Left" "Right"

> C # 内容を確認する

  Left Right
a    1     5
b    2     6
c    3     7
d    4     8

> ### データ例 datasets::airquality (help(airquality))
> dim(airquality)      # 大きさを確認

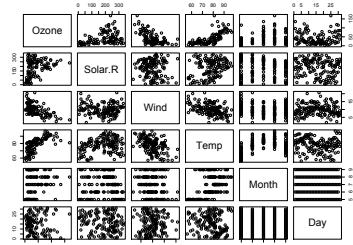
[1] 153   6

> names(airquality)    # 列の名前を表示
[1] "Ozone"   "Solar.R" "Wind"    "Temp"    "Month"
[6] "Day"

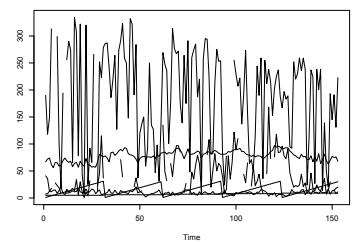
> head(airquality, n=5) # 最初の 5 つのデータを表示

  Ozone Solar.R Wind Temp Month Day
1    41      190  7.4   67     5    1
2    36      118  8.0   72     5    2
3    12      149 12.6   74     5    3
4    18      313 11.5   62     5    4
5    NA       NA 14.3   56     5    5
```

Rscript: `data-data.frame.r`



(a) 散布図



(b) 時系列のグラフ

図 3.4: データフレームの表示の例。

```

> plot(airquality)      # 散布図を表示
> ts.plot(airquality)   # 時系列として表示
> subset(airquality, Ozone>100) # 条件を満たす部分集合

  Ozone Solar.R Wind Temp Month Day
30     115     223  5.7   79      5  30
62     135     269  4.1   84      7  1
86     108     223  8.0   85      7 25
99     122     255  4.0   89      8  7
101    110     207  8.0   90      8  9
117    168     238  3.4   81      8 25
121    118     225  2.3   94      8 29

> subset(airquality, Ozone>100, select=Wind:Day)

  Wind Temp Month Day
30   5.7   79      5  30
62   4.1   84      7  1
86   8.0   85      7 25
99   4.0   89      8  7
101  8.0   90      8  9
117  3.4   81      8 25
121  2.3   94      8 29

```

実際のデータ解析においては、より複雑な操作を行う必要もあり、そうした操作に対応するためのパッケージも多数ある。中でも最近よく使われるパッケージ `dplyr` については補遺にて詳しく解説する。

演習 3.1. いろいろな形式のデータを作成してみよう。

- 関数 `c()`, `seq()`, `rep()`, `matrix()`, `array()`, `data.frame()`などの使い方を、関数 `help()` を用いて調べなさい。
- データの形式を調べる関数 `is.XXX` や、データの形式を変換する関数 `as.XXX` について調べなさい。
- 関数 `data()` を用いて、R にどのようなデータ集合が用意されているか調べなさい。

3.2 データの抽出

データから必要な部分集合を取り出すためには、添え字を指定するのが最も基本的な方法である。添え字の指定の仕方には、番号を指定する以外に、論理値で指定する方法がある。この場合、`TRUE` は要素の「選択」を、`FALSE` は要素の「除外」を意味する。また、前にも述べたように、要素に名前が付けられている場合は、その名前によってアクセス可能である。また、マイナス記号をつけて添え字番号を指定すると、その添え字番号の要素を除外する。

Rscript: `data-select.r`

```

> ### ベクトルの要素の指定
> x <- c(4, 1, 2, 9, 8, 3, 6)
> x[c(5,2)]      # 5番目と2番目の要素をこの順で抽出
[1] 8 1
> x[-c(2,3,7)]  # 2,3,7番目以外の要素を表示
[1] 4 9 8 3

```

```

> (idx <- x>3) # 3より大きい要素は TRUE, 3以下は FALSE
[1] TRUE FALSE FALSE TRUE TRUE FALSE TRUE
> which(idx) # TRUEに対応する要素の番号を表示
[1] 1 4 5 7
> x[idx] # 3より大きい要素(TRUE)をすべて表示
[1] 4 9 8 6
> x[x>3] # 上と同じ
[1] 4 9 8 6
> x[which(idx)] # 上と同じ
[1] 4 9 8 6
> x[-c(2,3,7)] # 2,3,7番目以外の要素を表示
[1] 4 9 8 3
> x[c(2,5)] <- c(0,1) # 2番目と5番目の要素を0と1に置換
> x
[1] 4 0 2 9 1 3 6
> names(x) <- letters[1:length(x)]
> ## xの要素にアルファベットを順に名前をつける
> x # 名前と内容を確認
a b c d e f g
4 0 2 9 1 3 6
> x[c("b","e")] # 2番目と5番目の要素
b e
0 1
> ### データフレームの要素の指定
> ### データ例 datasets::airquality (help(airquality))
> ### データにNA(欠損)があるので注意
> dim(airquality) # 大きさを確認
[1] 153 6
> names(airquality) # 列名を表示
[1] "Ozone" "Solar.R" "Wind" "Temp" "Month"
[6] "Day"
> head(airquality) # 最初の6行を表示
   Ozone Solar.R Wind Temp Month Day
1     41      190  7.4   67     5    1
2     36      118  8.0   72     5    2
3     12      149 12.6   74     5    3
4     18      313 11.5   62     5    4
5     NA       NA 14.3   56     5    5
6     28      NA 14.9   66     5    6
> str(airquality) # オブジェクトの構造を表示

```

```

'data.frame':      153 obs. of  6 variables:
 $ Ozone : int  41 36 12 18 NA 28 23 19 8 NA ...
 $ Solar.R: int  190 118 149 313 NA NA 299 99 19 194 ...
 $ Wind   : num  7.4 8 12.6 11.5 14.3 14.9 8.6 13.8 20.1 8.6 ...
 $ Temp   : int  67 72 74 62 56 66 65 59 61 69 ...
 $ Month  : int  5 5 5 5 5 5 5 5 5 5 ...
 $ Day    : int  1 2 3 4 5 6 7 8 9 10 ...

> airquality$Ozone>100 # Ozone が 100 を超えるかどうか

[1] FALSE FALSE FALSE FALSE NA FALSE FALSE FALSE
[9] FALSE NA FALSE FALSE FALSE FALSE FALSE FALSE
[17] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[25] NA NA NA FALSE FALSE TRUE FALSE NA
[33] NA NA NA NA NA FALSE NA FALSE
[41] FALSE NA NA FALSE NA NA FALSE FALSE
[49] FALSE FALSE FALSE NA NA NA NA NA
[57] NA NA NA NA NA TRUE FALSE FALSE
[65] NA FALSE FALSE FALSE FALSE FALSE FALSE NA
[73] FALSE FALSE NA FALSE FALSE FALSE FALSE FALSE
[81] FALSE FALSE NA NA FALSE TRUE FALSE FALSE
[89] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[97] FALSE FALSE TRUE FALSE TRUE NA NA FALSE
[105] FALSE FALSE NA FALSE FALSE FALSE FALSE FALSE
[113] FALSE FALSE NA FALSE TRUE FALSE NA FALSE
[121] TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[129] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[137] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[145] FALSE FALSE FALSE FALSE FALSE NA FALSE FALSE
[153] FALSE

> which(airquality$Ozone>100)

[1] 30 62 86 99 101 117 121

> ## Ozone が 100 を超える行の番号を抽出
> which(airquality$Ozone>100 & airquality$Wind<=5)

[1] 62 99 117 121

> ## 複数の条件の AND
> which(with(airquality, Ozone>100 & Wind<=5)) # 同上

[1] 62 99 117 121

> which(with(airquality, Ozone>100 | Wind<=5)) # OR

[1] 30 53 54 62 66 86 98 99 101 117 121 126 127

> airquality[which(airquality$Ozone>100), ] # 行の抽出

  Ozone Solar.R Wind Temp Month Day
30     115     223  5.7   79      5 30
62     135     269  4.1   84      7  1
86     108     223  8.0   85      7 25
99     122     255  4.0   89      8  7
101    110     207  8.0   90      8  9
117    168     238  3.4   81      8 25
121    118     225  2.3   94      8 29

> ## airquality[airquality$Ozone>100, ] # NA のため不可
> airquality[which(airquality$Ozone>100),
+             c("Month", "Day")] # 特定の列のみ表示

```

	Month	Day
30	5	30
62	7	1
86	7	25
99	8	7
101	8	9
117	8	25
121	8	29

データフレームから必要な部分集合を取り出す際に複雑な条件を指定する場合、添え字を指定するのではコードが読みにくくなってしまう。そのような場合にも対応できるように関数 `subset()` が用意されている。関数 `subset()` の基本的な書式は

```
subset(x, subset, select, drop=FALSE)
```

である。`x` にデータフレームを指定し、`subset` に抽出したい行に関する条件を、`select` に抽出したい列に関する条件をそれぞれ指定し、`drop` は結果が1行または1列のデータフレームになる場合にベクトルとして返すか否かを指定するオプションである。

```
> ### 関数 subset の使い方
> ### subset(dataframe, subset=条件, select=列)
> ## Ozone が 100 を超える行を抽出
> subset(airquality, subset = Ozone>100)

Ozone Solar.R Wind Temp Month Day
30     115     223  5.7    79      5   30
62     135     269  4.1    84      7   1
86     108     223  8.0    85      7   25
99     122     255  4.0    89      8   7
101    110     207  8.0    90      8   9
117    168     238  3.4    81      8   25
121    118     225  2.3    94      8   29

> ## Ozone が 100 を超える行で列名が Wind と Day
> subset(airquality, Ozone>100, select=c(Wind,Day))

Wind Day
30   5.7  30
62   4.1  1
86   8.0  25
99   4.0  7
101  8.0  9
117  3.4  25
121  2.3  29

> ## Ozone が 100 を超える行で列名が Wind から Day まで
> subset(airquality, Ozone>100, select=Wind:Day)

Wind Temp Month Day
30   5.7   79      5   30
62   4.1   84      7   1
86   8.0   85      7   25
99   4.0   89      8   7
101  8.0   90      8   9
117  3.4   81      8   25
121  2.3   94      8   29

> ## Ozone に欠測 (NA) がなく、かつ Day が 1 か 2 (AND)
> subset(airquality, !is.na(Ozone) & Day %in% c(1, 2))
```

Rscript: [data-subset.r](#)

```

      Ozone Solar.R Wind Temp Month Day
1       41      190  7.4   67      5    1
2       36      118  8.0   72      5    2
62      135      269  4.1   84      7    1
63      49      248  9.2   85      7    2
93      39      83   6.9   81      8    1
94       9      24  13.8   81      8    2
124     96      167  6.9   91      9    1
125     78      197  5.1   92      9    2

> ## Ozone が 100 以上か、または Wind が 5 以下 (OR)
> subset(airquality, Ozone > 100 | Wind <= 5)

      Ozone Solar.R Wind Temp Month Day
30      115      223  5.7   79      5   30
53       NA      59  1.7   76      6   22
54       NA      91  4.6   76      6   23
62      135      269  4.1   84      7    1
66      64      175  4.6   83      7    5
86      108      223  8.0   85      7   25
98      66       NA  4.6   87      8    6
99      122      255  4.0   89      8    7
101     110      207  8.0   90      8    9
117     168      238  3.4   81      8   25
121     118      225  2.3   94      8   29
126      73      183  2.8   93      9    3
127      91      189  4.6   93      9    4

> ## Day が 1 の行について、Temp 以外の列を抽出
> subset(airquality, Day == 1, select = -Temp)

      Ozone Solar.R Month Day
1       41      190      5    1
32      NA      286      6    1
62      135      269      7    1
93      39      83      8    1
124     96      167      9    1

```

その他、データフレームを特定のグループ分けに基づいて分割・結合するための関数 `split()`, `merge()` が用意されている(詳しい使い方はヘルプを参照)。また、より高度なデータフレームの加工を実行するための関数群がパッケージ `dplyr` に用意されている。

演習 3.2. データセット `datasets::airquality` (1973 年 5 月から 9 月までのニューヨークの大気の状態に関するデータ) から以下の条件を満たすデータを取り出せ。

1. 7 月のオゾン濃度 (`Ozone`)
2. 日射量 (`Solar.R`) に欠測 (NA) がないデータの月 (`Month`) と日 (`Day`)
3. 風速 (`Wind`) が時速 10 マイル以上で、気温 (`Temp`) が華氏 80 度以上の日のデータ

3.3 ファイルを用いたデータの読み書き

実際の解析の過程においては、収集されたデータを読み込んだり、整理したデータを保存したりする必要が生じる。R では一般に用

いられる CSV 形式 (comma separated values) のテキストファイルと, R の内部表現を用いたバイナリーファイル (ここでは RData 形式と呼ぶ) をサポートしている。以下では, データフレームを対象として, それぞれの形式でファイルの読み書きを行うための関数を纏める。

3.3.1 作業ディレクトリの確認と変更

R の実行は特定のフォルダ (ディレクトリ) 上で行われており, そのフォルダを **作業ディレクトリ** と呼ぶ。R のコード内でファイル名を指定した場合, 特に指定しない限り作業ディレクトリに存在するものとして扱われる。現在の作業ディレクトリは関数 `getwd()` で確認できる。作業ディレクトリの変更には関数 `setwd()` を利用するか, **RStudio** 上部の「Session」という項目から「Set Working Directory」を選び, その中の「Choose Directory...」という項目を選択すれば変更する作業ディレクトリを指定することができる。

```
> ### 作業ディレクトリの確認
> ## (環境によって実行結果が異なるため, 実行結果は省略)
> getwd()
> ### 作業ディレクトリの移動
> ## (環境によって指定の仕方も異なるので注意)
> setwd("~/Documents")
> ## ホームディレクトリ (~) 下の "Documents" フォルダに移動
```

Rscript: `data-wd.r`

3.3.2 CSV 形式の操作

1 つのデータフレームを CSV 形式のファイルへ書き出すには関数 `write.csv()` を用いる。書き出し後のファイルは特に指定しない限り作業ディレクトリ下に保存される。

```
> ### 関数 write.csv の使い方
> (myData <- subset(airquality, # データフレームの作成
+ Ozone>90, select=-Temp))

  Ozone Solar.R Wind Month Day
30     115     223   5.7     5   30
62     135     269   4.1     7    1
69      97     267   6.3     7    8
70      97     272   5.7     7    9
86     108     223   8.0     7   25
99     122     255   4.0     8    7
101    110     207   8.0     8    9
117    168     238   3.4     8   25
121    118     225   2.3     8   29
124     96     167   6.9     9    1
127     91     189   4.6     9    4

> dim(myData) # 大きさを確認
[1] 11  5
> write.csv(myData,file="myData.csv") # 書き出し
```

Rscript: `data-write.csv.r`

CSV 形式のファイルから読み込むには関数 `read.csv()` を用いる。読み込むファイルは, ディレクトリを明示的に指定しない限り, 作業ディレクトリに保存しておかなくてはいけない。

Rscript: data-read.csv.r

```

> ### 関数 read.csv の使い方
> (newdata <- read.csv(file="myData.csv", # 読み込み
+                         row.names=1)) # 1列目を行の名前

      Ozone Solar.R Wind Month Day
30     115     223  5.7    5   30
62     135     269  4.1    7   1
69      97     267  6.3    7   8
70      97     272  5.7    7   9
86     108     223  8.0    7  25
99     122     255  4.0    8   7
101    110     207  8.0    8   9
117    168     238  3.4    8  25
121    118     225  2.3    8  29
124     96     167  6.9    9   1
127     91     189  4.6    9   4

> dim(newdata) # 大きさを確認
[1] 11 5

> ### 東京都の気候データによる例 (tokyo_weather.csv)
> ## 気象庁のホームページより取得
> ## https://www.data.jma.go.jp/gmd/risk/obsdl/
> ## 地点・東京における平均気温(℃)・降水量(mm)など
> myData <- read.csv("data/tokyo_weather.csv",
+                      fileEncoding="utf8") # 文字コード
> head(myData) # データの最初の6行を表示

  年 月 日 曜日 気温 降水量 日射量 降雪量 風速
1 2018 1 1 月 6.2 0 11.59 0 2.7
2 2018 1 2 火 6.1 0 11.89 0 3.2
3 2018 1 3 水 4.9 0 11.77 0 5.2
4 2018 1 4 木 4.7 0 11.99 0 2.8
5 2018 1 5 金 3.7 0 4.07 0 1.8
6 2018 1 6 土 4.6 0 11.54 0 1.9

  最多風向 気圧 湿度 雲量 天気概況. 昼. 天気概況. 夜.
1 北北西 1009.7 55 1.0 晴 晴
2 北西 1011.2 42 0.3 快晴 晴
3 北西 1010.3 43 0.8 快晴 快晴
4 北北西 1015.0 41 0.8 快晴 晴後曇
5 北 1012.3 60 10.0 曇一時雨 晴一時曇
6 北西 1008.4 57 1.5 晴 晴

> dim(myData) # 大きさを確認
[1] 365 15

> colnames(myData) # 列名を確認
[1] "年"          "月"          "日"
[4] "曜日"        "気温"        "降水量"
[7] "日射量"      "降雪量"      "風速"
[10] "最多風向"    "気圧"        "湿度"
[13] "雲量"        "天気概況. 昼." "天気概況. 夜."

```

オプションとして与えられている `row.names=1` は、第1列を読み込んだデータフレームの各行の名前に割り当てる意味している。また日本語を含むファイルを読み込む場合は、ファイルを作成したアプリケーションやデータの配布元の情報に応じて、用いられている文字コードを指定する必要がある。良く用いられ

る文字コードとしては utf8 と sjis がある。

3.3.3 RData 形式の操作

RData 形式のファイルへの書き出しが関数 `save()` を用いる。関数 `write.csv()` と同様に、書き出し後のファイルは特に指定しない限り作業ディレクトリ下に保存される。CSV 形式と異なり、複数のデータフレームを 1 つのファイルに同時に保存することもできる。

```
> ### 関数 save の使い方
> (myDat1 <- subset(airquality, Temp>95, select=-Ozone))

  Solar.R Wind Temp Month Day
120      203   9.7    97     8   28
122      237   6.3    96     8   30

> (myDat2 <- subset(airquality, Temp<60, select=-Ozone))

  Solar.R Wind Temp Month Day
5        NA 14.3    56     5   5
8         99 13.8    59     5   8
15        65 13.2    58     5  15
18        78 18.4    57     5  18
21         8  9.7    59     5  21
25        66 16.6    57     5  25
26       266 14.9    58     5  26
27        NA  8.0    57     5  27

> dim(myDat1) # 大きさを確認
[1] 2 5
> dim(myDat2) # 大きさを確認
[1] 8 5
> save(myDat1,myDat2,file="mydata.rdata") # 書き出し
```

Rscript: `data-save.r`

RData 形式のファイルからの読み込みは関数 `load()` を用いる。関数 `read.csv()` と同様に、読み込むファイルはディレクトリを明示的に指定しない限り、作業ディレクトリに置かなくてはならない。

```
> ### 関数 load の使い方
> (myDat1 <- subset(airquality, Ozone > 120))

  Ozone Solar.R Wind Temp Month Day
62      135      269  4.1    84     7   1
99      122      255  4.0    89     8   7
117     168      238  3.4    81     8  25

> load(file="mydata.rdata") # 読み込み
> myDat1 # save したときの名前で読み込まれ上書きされる

  Solar.R Wind Temp Month Day
120      203   9.7    97     8   28
122      237   6.3    96     8   30

> myDat2

  Solar.R Wind Temp Month Day
5        NA 14.3    56     5   5
```

Rscript: `data-load.r`

8	99	13.8	59	5	8
15	65	13.2	58	5	15
18	78	18.4	57	5	18
21	8	9.7	59	5	21
25	66	16.6	57	5	25
26	266	14.9	58	5	26
27	NA	8.0	57	5	27

関数 `save()` では、データフレームの名前と内容が保存されるので、保存された名前が自動的に用いられる。したがって読み込む際には変数名の重複に注意が必要である。

演習 3.3. ファイル操作に慣れよう。

1. 関数 `write.csv()` で書き出したファイルの中身を適当なアプリケーションで確認しなさい。
2. 適当に作成したデータフレームをファイルに書き出しなさい。
3. 表を整理するには Excel などの表計算ソフトを用いるのが簡便であり、多くの表計算ソフトは CSV 形式でもデータが保存できるようになっている。自身の利用するソフトにおいて CSV 形式で保存する方法を調べなさい。
4. Excel 形式のファイルを直接読み込むパッケージもいくつかある。どのようなパッケージがあるか調べなさい。

3.4 データの整理

与えられたデータの総和や平均、最大値・最小値を求めたい状況は頻繁にある。**R** にはこれらの操作を簡便に実行するための関数としてそれぞれ `sum()`, `mean()`, `max()`, `min()` が用意されている。

Rscript: `data-summary.r`

```
> ### データの集計
> sum(1:100) # 1から 100までの整数の総和
[1] 5050

> ### 気候データによる例
> myData <- read.csv("data/tokyo_weather.csv",
+                      fileEncoding="utf8")
> temp <- myData$気温 # 気温を取り出す
> mean(temp) # 平均気温の計算
[1] 16.83973

> max(temp) # 最大値
[1] 32.2

> min(temp) # 最小値
[1] 0
```

行列もしくはデータフレームが与えられた際には、列(あるいは行)ごとに平均などの統計量を計算したい状況が頻繁にある。

そのような計算に便利な関数として関数 `apply()` がある。関数 `apply()` は基本的に以下のような書式で利用する:

```
apply(X, MARGIN, FUN)
```

ここで、引数 `X` にデータフレームを指定し、`MARGIN` には行ごとの計算には 1 を、列ごとの計算には 2 を指定する。引数 `FUN` には求めたい統計量を計算するための関数を指定する。

なお、総和や平均の場合には、列・行ごとに計算するための専用の関数が用意されており、それらを利用することもできる。

```
> ### 行・列ごとの計算
> x <- matrix(1:100, 4, 25)
> sum(x)           # x の成分の和を計算する (mean 等も同様)
[1] 5050

> rowSums(x)       # 行ごとの総和
[1] 1225 1250 1275 1300

> apply(x, 1, sum) # 上と同じ
[1] 1225 1250 1275 1300

> ### 気候データによる例
> myDat1 <- read.csv("data/tokyo_weather.csv",
+                      fileEncoding="utf8")
> myDat2 <- subset(myData, # 必要な列だけ選択
+                   select=c(気温, 降水量, 日射量, 風速))
> colMeans(myDat2)        # 列ごとの平均
    気温   降水量   日射量   風速
16.839726  3.960274 13.928740  2.954795

> apply(myDat2, 2, max) # 列ごとの最大値
    気温 降水量 日射量 風速
32.2   58.0   29.7   7.4

> sapply(myDat2, max)   # 上と同じ
    気温 降水量 日射量 風速
32.2   58.0   29.7   7.4

> apply(myDat2, 2, min) # 列ごとの最小値
    気温 降水量 日射量 風速
0.00   0.00   0.92   1.20

> ## 自作関数の適用
> ## 列ごとに平均より大きいデータ数 (TRUE=1/FALSE=0) を計算
> apply(myDat2, 2, function(x){sum(x>mean(x))})

    気温 降水量 日射量 風速
194      70     162    137
```

Rscript: `data-apply.r`

データフレームの各行をいくつかのグループにまとめて、グループごとの統計量を計算したい状況も頻繁に生じる。この場合に便利なのが関数 `aggregate()` である。関数 `aggregate()` は基本的に以下のような書式で利用する:

```
aggregate(X, BY, FUN)
```

ここで、引数 X にデータフレームを指定し、BY には各行が属するグループを指定するベクトルをリストで与える(複数可)。引数 FUN には求めたい統計量を計算するための関数を指定する。なお X がベクトルの場合には関数 tapply() も利用可能である。

Rscript: `data-aggregate.r`

```
> ### 気候データによる例
> myDat1 <- read.csv("data/tokyo_weather.csv",
+                      fileEncoding="utf8")
> myDat2 <- subset(myData,
+                   select=c(気温, 降水量, 日射量, 風速))
> myDat3 <- subset(myData,
+                   select=c(月, 気温, 降水量, 日射量, 風速))
> ## 月ごとの各変数の平均値を求める
> aggregate(myDat2,
+            by=list(月=myDat1$月), # 条件に合う行を選択
+            FUN=mean) # 各列に適用する関数

  月    気温    降水量    日射量    風速
1 1 4.654839 1.5645161 10.286452 2.638710
2 2 5.353571 0.7142857 12.147500 2.550000
3 3 11.461290 7.0967742 14.414516 3.012903
4 4 17.010000 3.6333333 17.507333 3.486667
5 5 19.803226 5.3387097 18.650645 3.270968
6 6 22.376667 5.1833333 16.897000 3.166667
7 7 28.254839 3.4516129 19.777097 3.519355
8 8 28.061290 2.7903226 18.361613 3.464516
9 9 22.856667 12.1666667 10.791667 3.006667
10 10 19.112903 1.9838710 10.872581 2.603226
11 11 13.973333 2.1000000 9.460333 2.140000
12 12 8.332258 1.4193548 7.771613 2.558065

> ## 以下のコードも同じ結果を返す
> aggregate(. ~ 月, data=myDat3, FUN=mean)

  月    気温    降水量    日射量    風速
1 1 4.654839 1.5645161 10.286452 2.638710
2 2 5.353571 0.7142857 12.147500 2.550000
3 3 11.461290 7.0967742 14.414516 3.012903
4 4 17.010000 3.6333333 17.507333 3.486667
5 5 19.803226 5.3387097 18.650645 3.270968
6 6 22.376667 5.1833333 16.897000 3.166667
7 7 28.254839 3.4516129 19.777097 3.519355
8 8 28.061290 2.7903226 18.361613 3.464516
9 9 22.856667 12.1666667 10.791667 3.006667
10 10 19.112903 1.9838710 10.872581 2.603226
11 11 13.973333 2.1000000 9.460333 2.140000
12 12 8.332258 1.4193548 7.771613 2.558065

> ## 月および降水の有無でグループ分け
> aggregate(myDat2, FUN=mean,
+            by=list(月=myDat3$月,
+                    降水の有無=(myDat3$降水量>0)))

  月 降水の有無    気温    降水量    日射量    風速
1 1 FALSE 4.507407 0.000000 11.236296 2.670370
2 2 FALSE 5.568182 0.000000 13.554545 2.577273
3 3 FALSE 12.336842 0.000000 18.645789 2.826316
4 4 FALSE 17.162500 0.000000 19.947500 3.462500
5 5 FALSE 20.871429 0.000000 22.742381 3.276190
6 6 FALSE 23.855556 0.000000 22.627222 3.577778
7 7 FALSE 28.829167 0.000000 21.814583 3.445833
8 8 FALSE 28.642857 0.000000 21.210952 3.495238
```

9	9	FALSE	24.236364	0.000000	15.609091	3.136364
10	10	FALSE	19.278947	0.000000	11.855789	2.589474
11	11	FALSE	13.657895	0.000000	11.116842	2.205263
12	12	FALSE	8.516667	0.000000	8.971250	2.645833
13	1	TRUE	5.650000	12.125000	3.875000	2.425000
14	2	TRUE	4.566667	3.333333	6.988333	2.450000
15	3	TRUE	10.075000	18.333333	7.715000	3.308333
16	4	TRUE	16.400000	18.166667	7.746667	3.583333
17	5	TRUE	17.560000	16.550000	10.058000	3.260000
18	6	TRUE	20.158333	12.958333	8.301667	2.550000
19	7	TRUE	26.285714	15.285714	12.791429	3.771429
20	8	TRUE	26.840000	8.650000	12.378000	3.400000
21	9	TRUE	22.057895	19.210526	8.002632	2.931579
22	10	TRUE	18.850000	5.125000	9.315833	2.625000
23	11	TRUE	14.518182	5.727273	6.599091	2.027273
24	12	TRUE	7.700000	6.285714	3.658571	2.257143

演習 3.4. R に含まれるデータセット `datasets::airquality` を用いてデータの整理をしてみよう。

1. 月日以外の変数ごとに平均、最大値および最小値を求めよ。
2. 月ごとの平均、最大値および最小値を求めよ。

3.5 補遺

3.5.1 参考文献

この章に関連する参考書としては以下を挙げておく。

- [1] 金明哲. *R によるデータサイエンス (第 2 版)*. 東京: 森北出版, 2017.
- [2] U. リゲス (石田基広訳). *R の基礎とプログラミング技法*. 東京: 丸善出版, 2012.
- [3] 奥村晴彦. *R で楽しむ統計*. 東京: 共立出版, 2016.
- [4] 山本義郎, 藤野友和, 久保田貴文. *R によるデータマイニング入門*. 東京: オーム社, 2015.

3.5.2 パッケージ `dplyr` によるデータの操作

先に紹介した関数 `subset()` を含め, R の基本パッケージにはデータを整理するための簡単な機能を提供する関数が用意されているので、これらの関数を組み合わせて希望の操作を行うことが可能である。しかしながら、複雑な操作を行う場合にはプログラムが繁雑となるので、より強力な関数群を集めたパッケージがいくつか用意されている。その1つとしてパッケージ `dplyr` がある。以下では、その基本的な使い方を紹介する。

まず、以下のようにしてパッケージ `dplyr` を読み込んでおく。

```
> library(dplyr) # パッケージ dplyr を読み込む
```

Rscript: `dplyr-require.r`

3.5.3 条件を指定した行の選択

特定の条件を満たす行を選択するには関数 `dplyr::filter()` を用いる。単に行の番号を指定して選択するには関数 `dplyr::slice()` を用いればよい。

Rscript: `dplyr-filter.r`

```
> ### dplyr::filter (条件で絞り込む)
> filter(airquality, Temp>80, Ozone>90) # 条件の and

  Ozone Solar.R Wind Temp Month Day
1    135      269  4.1   84     7    1
2     97      267  6.3   92     7    8
3     97      272  5.7   92     7    9
4    108      223  8.0   85     7   25
5    122      255  4.0   89     8    7
6    110      207  8.0   90     8    9
7    168      238  3.4   81     8   25
8    118      225  2.3   94     8   29
9     96      167  6.9   91     9    1
10    91      189  4.6   93     9    4

> filter(airquality, Day==5 | Day==10) # 条件の or

  Ozone Solar.R Wind Temp Month Day
1     NA       NA 14.3   56     5    5
2     NA      194  8.6   69     5   10
3     NA      220  8.6   85     6    5
4     39      323 11.5   87     6   10
5     64      175  4.6   83     7    5
6     85      175  7.4   89     7   10
7     35       NA  7.4   85     8    5
8     NA      222  8.6   92     8   10
9     47       95  7.4   87     9    5
10    24      259  9.7   73     9   10

> ### dplyr::slice (行番号で絞り込む)
> slice(airquality, seq(2,16,by=2))

  Ozone Solar.R Wind Temp Month Day
1     36      118  8.0   72     5    2
2     18      313 11.5   62     5    4
3     28       NA 14.9   66     5    6
4     19      99 13.8   59     5    8
5     NA      194  8.6   69     5   10
6     16      256  9.7   69     5   12
7     14      274 10.9   68     5   14
8     14      334 11.5   64     5   16
```

3.5.4 列の値による行の並べ替え

特定の列に入っている値の昇順、あるいは降順に並べ替えるには関数 `dplyr::arrange()` を用いる。特に指定がなければ昇順で並べ替えが行われるが、降順で並べ替えたい場合には関数 `dplyr::desc()` の中に入れて列を指定すればよい。また、順次操作を行う操作を簡略化するために演算子 `%>%` が用意されている。

Rscript: `dplyr-arrange.r`

```
> ### dplyr::arrange (指定した変数の順に並べ替える)
> arrange(airquality, Temp, Wind) %>% # 結果を次に渡す
+     head(8) # 先頭の 8 行を表示。並べ替えの基本は昇順
```

```

Ozone Solar.R Wind Temp Month Day
1   NA      NA 14.3  56      5   5
2   NA      NA  8.0  57      5  27
3   NA      66 16.6  57      5  25
4    6      78 18.4  57      5  18
5   18      65 13.2  58      5  15
6   NA     266 14.9  58      5  26
7    1       8  9.7   59      5  21
8   19      99 13.8  59      5   8

> arrange(airquality, Temp, desc(Wind)) %>%
+   head(8) # Wind を降順に指定

Ozone Solar.R Wind Temp Month Day
1   NA      NA 14.3  56      5   5
2    6      78 18.4  57      5  18
3   NA      66 16.6  57      5  25
4   NA      NA  8.0  57      5  27
5   NA     266 14.9  58      5  26
6   18      65 13.2  58      5  15
7   19      99 13.8  59      5   8
8    1       8  9.7   59      5  21

```

3.5.5 列の選択

特定の列または列の集合を選択するには関数 `dplyr::select()` を用いる。列の指定に仕方には、必要とする列を指定する方法と、不要な列を除く方法がある。また、選択と同時に列名を変更することもできる。なお、選択せずに名前のみを変更するには関数 `dplyr::rename()` を用いる。

```

> ### dplyr::select (列を選択する)
> select(airquality, Temp, Wind, Ozone) %>%
+   head(8) # 列記する場合

Temp Wind Ozone
1   67  7.4   41
2   72  8.0   36
3   74 12.6   12
4   62 11.5   18
5   56 14.3    NA
6   66 14.9   28
7   65  8.6   23
8   59 13.8   19

> select(airquality, Ozone:Temp) %>%
+   head(8) # 最初と最後の列を指定する場合

Ozone Solar.R Wind Temp
1    41      190  7.4   67
2    36      118  8.0   72
3    12      149 12.6   74
4    18      313 11.5   62
5    NA      NA 14.3   56
6    28      NA 14.9   66
7    23      299  8.6   65
8    19      99 13.8   59

> select(airquality, -(Month:Day)) %>%
+   head(8) # 削除する場合

```

Rscript: `dplyr-select.r`

```

      Ozone Solar.R Wind Temp
1       41      190  7.4   67
2       36      118  8.0   72
3       12      149 12.6   74
4       18      313 11.5   62
5       NA      NA 14.3   56
6       28      NA 14.9   66
7       23      299  8.6   65
8       19      99 13.8   59

> select(airquality, TempInF=Temp) %>%
+     head(8) # 名前を変更する場合

      TempInF
1       67
2       72
3       74
4       62
5       56
6       66
7       65
8       59

> rename(airquality, TempInF=Temp) %>%
+     head(8) # 指定箇所だけ名前を変更する場合

      Ozone Solar.R Wind TempInF Month Day
1       41      190  7.4      67      5   1
2       36      118  8.0      72      5   2
3       12      149 12.6      74      5   3
4       18      313 11.5      62      5   4
5       NA      NA 14.3      56      5   5
6       28      NA 14.9      66      5   6
7       23      299  8.6      65      5   7
8       19      99 13.8      59      5   8

```

3.5.6 値の整理

関数 `dplyr::distinct()` を用いると、各列でどのような値が使われているかを知ることができる。こうした集計は特にカテゴリカル変数の場合に重要となり、複数の列での組み合わせにも対応している。

Rscript: `dplyr-distinct.r`

```

> ### dplyr::distinct (異なる値を取り出す)
> distinct(airquality, Month) # どの月が対象か調べる

      Month
1       5
2       6
3       7
4       8
5       9

> distinct(airquality, Ozone, Temp) %>%
+     head(8) # Ozone と Temp の異なる組み合わせ

      Ozone Temp
1       41    67
2       36    72
3       12    74
4       18    62

```

```

5     NA   56
6     28   66
7     23   65
8     19   59

> ### dplyr::count (異なる値がいくつあるか数える)
> count(airquality, Month) # 各月のデータ数を調べる

# A tibble: 5 x 2
  Month     n
  <int> <int>
1     5     31
2     6     30
3     7     31
4     8     31
5     9     30

```

3.5.7 列の追加

新しい量を計算して、新たな列を加えるには関数 `dplyr::mutate()` を用いる。新たな列のみで他の列を保持する必要がない場合には関数 `dplyr::transmute()` を用いればよい。

```

> ### dplyr::mutate (新しい列を作成する)
> mutate(airquality, TempC=(Temp-32)/1.8) %>%
+     head(8) # 華氏を摂氏に変換して追加

  Ozone Solar.R Wind Temp Month Day    TempC
1     41      190  7.4   67      5   1 19.44444
2     36      118  8.0   72      5   2 22.22222
3     12      149 12.6   74      5   3 23.33333
4     18      313 11.5   62      5   4 16.66667
5     NA       NA 14.3   56      5   5 13.33333
6     28       NA 14.9   66      5   6 18.88889
7     23      299  8.6   65      5   7 18.33333
8     19      99 13.8   59      5   8 15.00000

> ### dplyr::transmute (新しい列を作成する)
> transmute(airquality,
+             TempF=Temp, TempC=(Temp-32)/1.8) %>%
+     head(8) # 元の列を保持しない

  TempF    TempC
1     67 19.44444
2     72 22.22222
3     74 23.33333
4     62 16.66667
5     56 13.33333
6     66 18.88889
7     65 18.33333
8     59 15.00000

```

Rscript: [dplyr-mutate.r](#)

3.5.8 データフレームの集計

データフレームの各行での集計値を得るには関数 `dplyr::summarize()` を用いる。どのような集計を行うかは適切な関数を指定する必要がある。

Rscript: dplyr-summarize.r

```
> ### dplyr::summarize (データフレームを集計する)
> summarize(airquality,
+           mean = mean(Ozone, na.rm=TRUE),
+           min = min(Ozone, na.rm=TRUE),
+           median = median(Ozone, na.rm=TRUE),
+           max = max(Ozone, na.rm=TRUE))

  mean min median max
1 42.12931   1    31.5 168
```

3.5.9 行のリサンプリング

データフレームの中からランダムに行をリサンプリングするには関数 `dplyr::sample_n()` または `dplyr::sample_frac()` を用いる。関数 `dplyr::sample_n()` は全体の中から指定した個数を、関数 `dplyr::sample_frac()` は全体の中の指定して割合をリサンプリングする。これらはブートストラップ法 (bootstrap method) や交差検証法 (cross-validation method) などに利用される。

Rscript: dplyr-sample.r

```
> ### dplyr::sample_n (個数を指定してサンプリングする)
> sample_n(airquality, size=10)

  Ozone Solar.R Wind Temp Month Day
1     52       82  12.0   86      7   27
2     NA       59   1.7   76      6   22
3     NA      220   8.6   85      6   5
4     59      254   9.2   81      7  31
5     10      264  14.3   73      7  12
6     71      291  13.8   90      6   9
7     37      284  20.7   72      6  17
8     39       83   6.9   81      8   1
9     16      201   8.0   82      9  20
10    110     207   8.0   90      8   9

> ### dplyr::sample_frac (比率を指定してサンプリングする)
> sample_frac(airquality, size=0.05)

  Ozone Solar.R Wind Temp Month Day
1     NA      322  11.5   79      6  15
2     23      299   8.6   65      5   7
3     96      167   6.9   91      9   1
4     32      236   9.2   81      7   3
5     30      193   6.9   70      9  26
6     39       83   6.9   81      8   1
7     50      275   7.4   86      7  29
8     NA      291  14.9   91      7  14
```

なお、復元抽出を行う場合には関数 `sample()` と同様にオプション `replace=TRUE` を付ければよい。

3.5.10 データフレームのグループ化

関数 `dplyr::group_by()` を用いると、特定の行でグループ化して、その集計を求めることができる。この関数は、上記の関数と組み合わせ、集計途中を保存しながら次の関数に引き渡すようにして用いることが多い。

```

> ### dplyr::group_by (指定した変数でグループ化する)
> air_month <- group_by(airquality, Month) # 月ごと
> summarize(air_month,
+           count = n(), # 各月の日数を求める
+           mean = mean(Ozone, na.rm=TRUE),
+           min = min(Ozone, na.rm=TRUE),
+           median = median(Ozone, na.rm=TRUE),
+           max = max(Ozone, na.rm=TRUE))

# A tibble: 5 x 6
  Month count  mean   min median   max
  <int> <int> <dbl> <int> <dbl> <int>
1     5     31 23.6     1    18   115
2     6     30 29.4    12    23    71
3     7     31 59.1     7    60   135
4     8     31 60.0     9    52   168
5     9     30 31.4     7    23    96

> ## 演算子%>%を用いることで変数を用いずに簡略化できる
> airquality %>%
+   group_by(Month) %>%
+   summarize(
+     count = n(),
+     mean = mean(Ozone, na.rm=TRUE),
+     min = min(Ozone, na.rm=TRUE),
+     median = median(Ozone, na.rm=TRUE),
+     max = max(Ozone, na.rm=TRUE))

# A tibble: 5 x 6
  Month count  mean   min median   max
  <int> <int> <dbl> <int> <dbl> <int>
1     5     31 23.6     1    18   115
2     6     30 29.4    12    23    71
3     7     31 59.1     7    60   135
4     8     31 60.0     9    52   168
5     9     30 31.4     7    23    96

```

Rscript: [dplyr-groupby.r](#)

演習 3.5. データフレームを整理してみよう。

1. 関数 `data()` で調べた適当なデータフレームを整理しなさい。
2. 開発者たちによる解説は `vignette("dbplyr")` で読むことができる。パッケージ `nycflights13` をインストールして例を確認しなさい。

3.5.11 その他

CSV 形式でない通常のテキストファイルを読み込むための関数として `read.table()` がある。ファイルの大きさが大きい場合、読み込みや書き出しに非常に時間がかかることがある。その場合、ファイル操作を高速化したパッケージ群がいくつか開発されているため、それらを利用するのが便利である。例えば、大規模データの読み込みにはパッケージ `data.table` の関数 `fread()` が便利である。CSV ファイルの書き出しにはパッケージ `readr` の関数 `write_csv()` が便利である。

データのプロット

記述統計量と並んでデータ全体の特徴や傾向を把握するために効果的な方法は、データを可視化することである。R の基本パッケージ `graphics` に用意されている作図機能はきわめて多彩であり、これらを適切に組み合わせることによって様々な種類のグラフを描くことができる。以下では、いくつかの代表的な描画関数を取り上げて解説する。

描画関連の関数は色、線種や線の太さ、あるいは図中の文字の大きさなどを指定するために、多彩なオプションを用意しているので、必要に応じて関数 `help()`(ヘルプの表示) と `example()`(例題の表示) を利用して欲しい。

4.1 基本的な描画

描画において基本となるのは関数 `plot()` である。

関数 `sin()` のように 1 変数の関数として定義されているものは、定義域を指定してやればそのまま表示することができる。関数を追加するにはオプション `add` とともに関数 `curve()` を用いれば良い。

また、関数 `plot()` に同じ長さの二つのベクトルを与えると、同じ番号の要素からなる点の組 (x, y) をプロットして、その**散布図**を描くことができる。

プロットの種類(点や線)を指定するにはオプション `type` を用いる。`'p'` で点(point), `'l'` で点列を順に結んだ線(line)が描かれる。なお、オプションに与える文字列は'(シングルクオート)か"(ダブルクオート)で囲む必要がある。

オプション `col` で”色の名前”を指定することにより点や線の色を変えることができる。R で指定することのできる色の名前は関数 `colors()` で照会することができる。

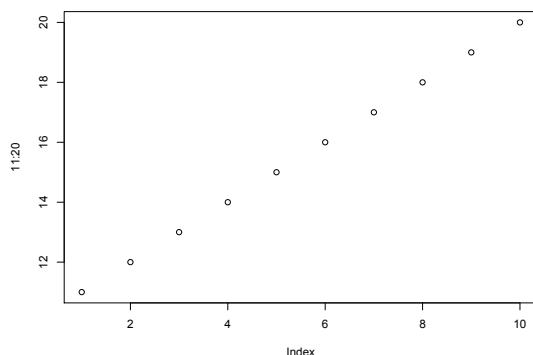
関数 `plot()` で描いた図中に更に線を追加するには関数 `lines()` を、点を追加するには関数 `points()` を用いる。

これ以外にも関数 `plot()` は様々なオプションを指定することができる。 `help(plot)` および `help(plot.default)` を参照して欲しい。

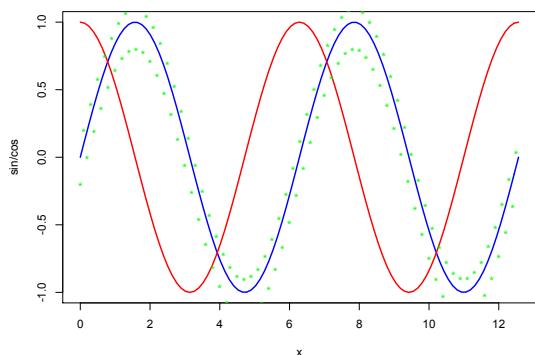
図 4.1 参照

```
> ### ベクトルの描画
> plot(11:20)
> ### 関数の描画
> plot(sin, 0, 4*pi,
+       col="blue", # グラフの線の色
+       lwd=2, # グラフの線の太さ
+       ylab="sin/cos") # y 軸のラベル
> curve(cos,
+        add=TRUE, # グラフを上書き
+        col="red", lwd=2)
> x <- seq(0, 4*pi, by=0.1)
> y <- sin(x) + rep_len(c(-0.2, 0.1), length(x))
```

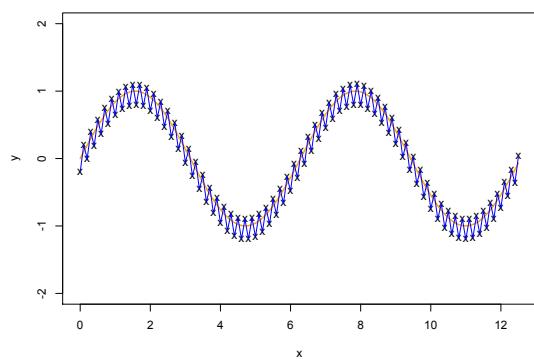
Rscript: `graph-plot.r`



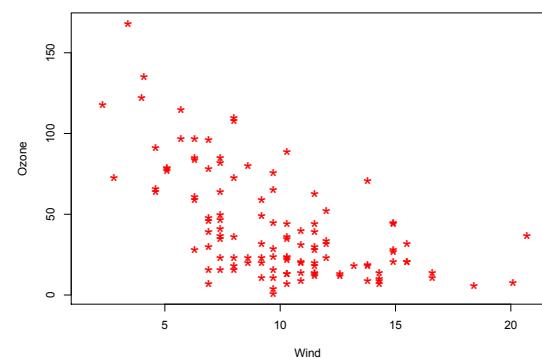
(a) ベクトルのプロット



(b) 関数の描画



(c) データの描画



(d) 散布図

図 4.1: 関数 `plot()` の例.

```
> points(x, y, col="green", # 点を追加
+         pch="*") # pch は点の形を指定
> ### データ点の描画
> plot(x, y, type="p", pch="x", # "p"=point
+       ylim=c(-2,2)) # ylim で値域を指定
> curve(sin, add=TRUE, col="orange", lwd=2)
> lines(x, y, col="blue") # 折れ線を追加
> ### データフレームを用いた散布図 (airquality を利用)
> plot(Ozone ~ Wind, data=airquality,
+       pch="*", col="red",
+       cex=2) # cex は点の大きさの倍率を指定
```

関数 `legend()` によってグラフに凡例を追加することができる。また、以下の例で見るよう **R** には数式を扱う機能がある。詳細は `help(plotmath)` を参照してほしい。

Rscript: `graph-legend.r`

図 4.2 参照

```
> ### 凡例の追加
> f <- function(x) exp(-x) * cos(x)
> plot(f, 0, 2*pi, col="red", lwd=2, ylab="")
> g <- function(x) exp(-x) * sin(x)
> curve(g, lty=2, # グラフの線の形式 2 は破線
+        add=TRUE, col="blue", lwd=2)
> legend(4, # 凡例の左上の x 座標
+        1, # 凡例の左上の y 座標
```

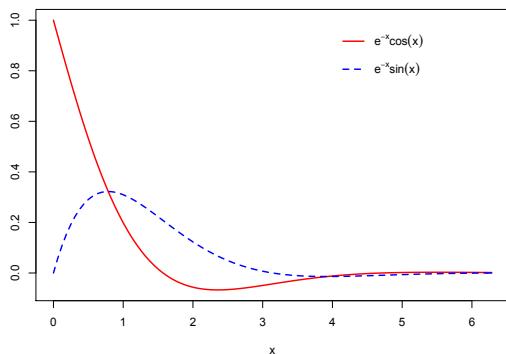


図 4.2: 関数 legend() の例.

```
+ legend=c(expression(e^{-x}*cos(x)),
+           expression(e^{-x}*sin(x))), # 凡例
+ lty=c(1,2), lwd=2,
+ col=c("red","blue"), # plot に準拠
+ bty="n", # 凡例の枠線の形式 ("n"は枠線なし)
+ y.intersp=2) # 行間の指定
```

なお、OSによっては日本語を含む図を描画すると文字化けする場合がある。その場合関数 par() のオプション family に適当なフォントファミリーを指定することで文字化けを回避できる。例えば Mac OS のデフォルトの設定では日本語を含む図は文字化けしてしまうが、以下のコマンドをコンソール上で実行することで文字化けを回避できる。

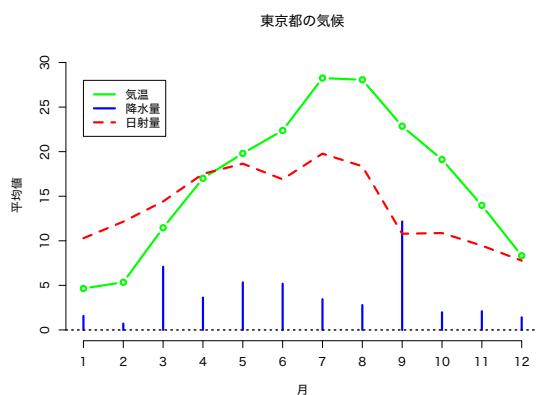


図 4.3: 日本語フォントを指定した例.

図 4.3 参照Rscript: [graph-font.r](#)

```
> ### 日本語フォントの指定
> par(family="HiraginoSans-W4")
> ### 東京都の気候データによる例 (tokyo_weather.csv)
> ## 気象庁のホームページより取得
> ## https://www.data.jma.go.jp/gmd/risk/obsdl/
> ## 地点・東京における平均気温 (°C)・降水量 (mm)など
> myData <- subset(read.csv("data/tokyo_weather.csv",
+                           fileEncoding="utf8"),
+                           select=c(月, 気温, 降水量, 日射量, 風速))
> ## 月ごとの平均をプロットする
> (x <- aggregate(. ~ 月, data=myData, FUN=mean))
```

	月	気温	降水量	日射量	風速
1	1	4.654839	1.5645161	10.286452	2.638710
2	2	5.353571	0.7142857	12.147500	2.550000
3	3	11.461290	7.0967742	14.414516	3.012903
4	4	17.010000	3.6333333	17.507333	3.486667
5	5	19.803226	5.3387097	18.650645	3.270968
6	6	22.376667	5.1833333	16.897000	3.166667
7	7	28.254839	3.4516129	19.777097	3.519355
8	8	28.061290	2.7903226	18.361613	3.464516
9	9	22.856667	12.1666667	10.791667	3.006667
10	10	19.112903	1.9838710	10.872581	2.603226
11	11	13.973333	2.1000000	9.460333	2.140000
12	12	8.332258	1.4193548	7.771613	2.558065

```

> plot(x$気温, type = "b", # 点と線 (both)
+       lwd=3, col="green", # 太さと色
+       ylim=c(0, max(x$気温)+2), # y軸の範囲を限定
+       xlab="月", ylab="平均値", main="東京都の気候",
+       axes=FALSE) # 軸を書かない
> axis(1, 1:12, 1:12) # x軸の作成
> axis(2) # y軸の作成
> lines(x$降水量, type="h", lwd=3, col="blue")
> lines(x$日射量, lwd=3, lty=2, col="red")
> abline(h=0, lwd=2, lty="dotted") # y=0の線を引く
> legend(1, 28, # 凡例の位置を左上の座標で指定
+         legend=c("気温", "降水量", "日射量"),
+         col=c("green", "blue", "red"),
+         lty=c(1,1,2), lwd=3)

```

作成したグラフは保存することができる。RStudio の機能を使う場合には右下ペインの “Plots” タブの “Export” をクリックすると、形式やサイズを指定して保存できる（もしくはクリップボードにコピーもできる）。コマンドで実行することも可能であるが、それについての詳細は `help(png)` や `help(dev.copy)` を参照してほしい。

4.2 ヒストグラム

データの頻度分布を表すヒストグラムを描画するには関数 `hist()` を用いる。これ以外にも凝ったヒストグラムを書くための関数がいくつか用意されているが、これらについては `help.search("histogram")` を参照して欲しい。

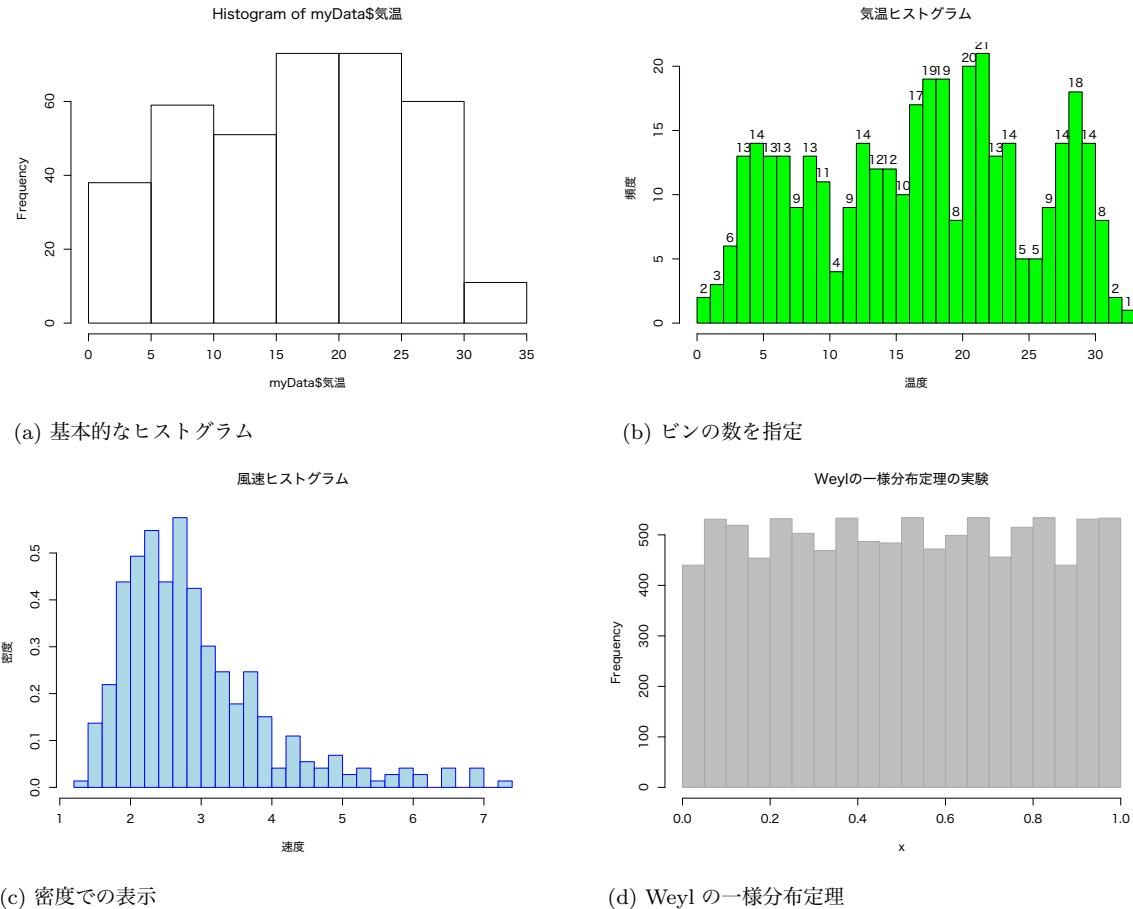
Rscript: `graph-hist.r`

図 4.4 参照

```

> ### 関数 hist によるヒストグラムの作図
> par(family="HiraginoSans-W4")
> myData <- subset(read.csv("data/tokyo_weather.csv",
+                           fileEncoding="utf8"),
+                           select=c(月, 気温, 降水量, 日射量, 風速))
> ### 基本的なヒストグラム
> hist(myData$気温) # ビンの数は自動的に計算される
> ### ビンの数を指定
> hist(myData$気温,
+       breaks=25, # ビンの数を約 25 に設定
+       labels=TRUE, # 各ビンの度数を表示
+       col="green",

```

図 4.4: 関数 `hist()` の例.

```
+      xlab="温度", ylab="頻度", main="気温ヒストグラム")
> ### 密度での表示
> hist(myData$風速, freq = FALSE, # 全体に対する割合で表示
+       breaks=25, col="lightblue",
+       border="blue", # 長方形の境界の色
+       xlab="速度", ylab="密度", main="風速ヒストグラム")
> ### Weyl の一様分布定理の確認
> ## aが無理数のとき、数列 a, 2a, 3a, ... の小数部分は
> ## 区間 (0,1) 上に均一に現れる
> a <- pi # 無理数 (適当に変えて実験してみよう)
> n <- 10000
> x <- (1:n) * a
> x <- x - floor(x) # 小数の取り出し (floor は Gauss 記号)
> hist(x, breaks=20, col="gray", border="darkgray",
+       main="Weyl の一様分布定理の実験")
```

4.3 箱ひげ図

複数のデータの分布を比較する際、観測数が大きく異なるなどヒストグラムでの比較が難しい場合がある。複数のデータの分布の違いを見るには箱ひげ図(boxplot)が良く用いられるが、これは関数 `boxplot()` で描くことができる。

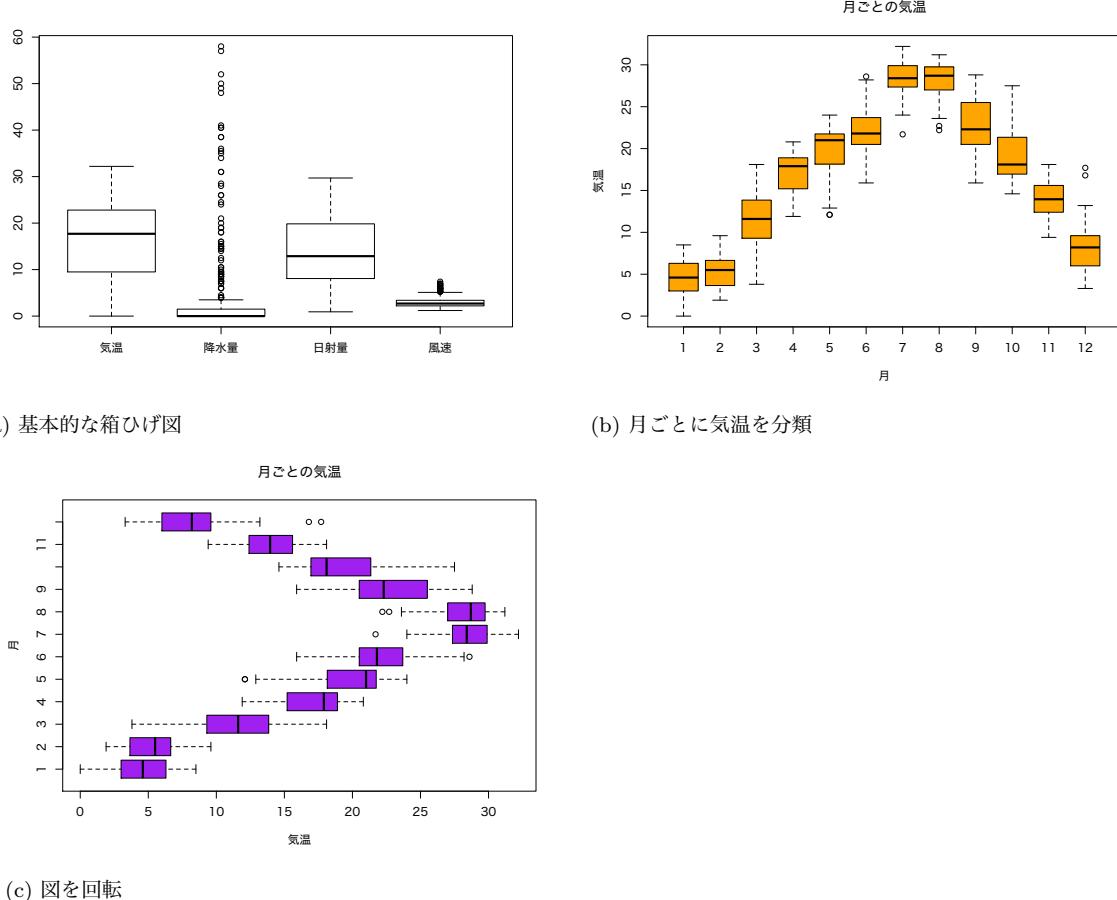


図 4.5: 関数 boxplot() の例.

Rscript: [graph-boxplot.r](#)**図 4.5 参照**

```
> ### 関数 boxplot による箱ひげ図の作図
> par(family="HiraginoSans-W4")
> myData <- subset(read.csv("data/tokyo_weather.csv",
+                           fileEncoding="utf8"),
+                           select=c(月, 気温, 降水量, 日射量, 風速))
> ### 基本的な箱ひげ図
> boxplot(myData[, -1]) # 月は除く
> ### 月ごとに気温を分類
> boxplot(気温 ~ 月, data=myData,
+           col="orange", main="月ごとの気温")
> boxplot(気温 ~ 月, data=myData,
+           horizontal=TRUE, # 図を回転
+           col="purple", main="月ごとの気温")
```

4.4 棒グラフ

関数 barplot() によって棒グラフを作成できる。barplot() の第1引数はベクトルまたは行列でなければならないことに注意する。

Rscript: [graph-barplot.r](#)**図 4.6 参照**

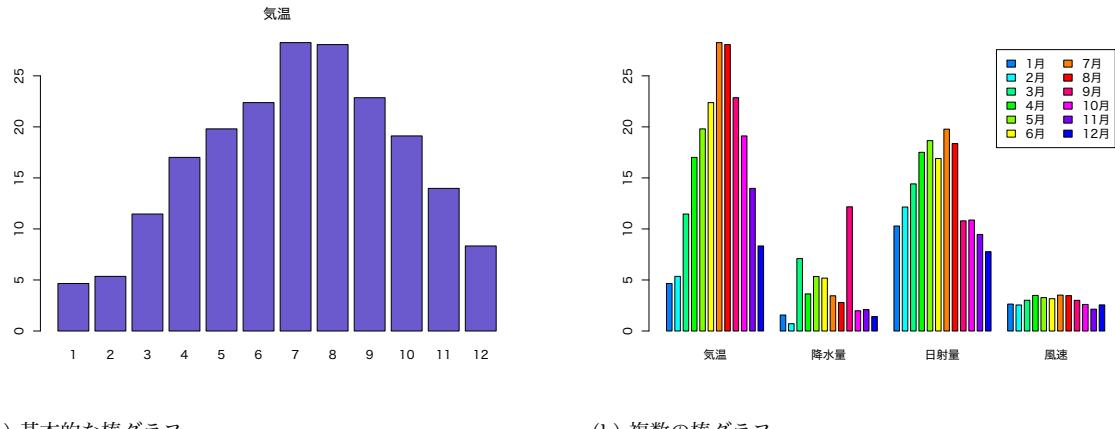


図 4.6: 関数 barplot() の例.

```
> ### 関数 barplot による棒グラフの作図
> par(family="HiraginoSans-W4")
> myData <- subset(read.csv("data/tokyo_weather.csv",
+                         fileEncoding="utf8"),
+                     select=c(月, 気温, 降水量, 日射量, 風速))
> x <- aggregate(. ~ 月, data=myData, FUN=mean)
> ### 基本的な棒グラフ
> barplot(x[,2], # 棒の高さのベクトル
+           col="slateblue", # 棒の色の指定
+           names.arg=x[,1], # x軸のラベル
+           main=names(x)[2]) # タイトル
> ### 複数の棒グラフ
> barplot(as.matrix(x[,-1]), # 第1引数はベクトル/行列
+           col=rainbow(12)[c(8:1,12:9)], # 12色に色分け
+           beside=TRUE, # 棒グラフを横に並べる
+           space=c(.5, 3), # 棒間・変数間のスペースを指定
+           legend.text=paste0(x[,1], "月"), # 凡例の指定
+           args.legend=list(ncol=2)) # 凡例を2列にして表示
```

4.5 円グラフ

円グラフは関数 pie() で描くことができる。

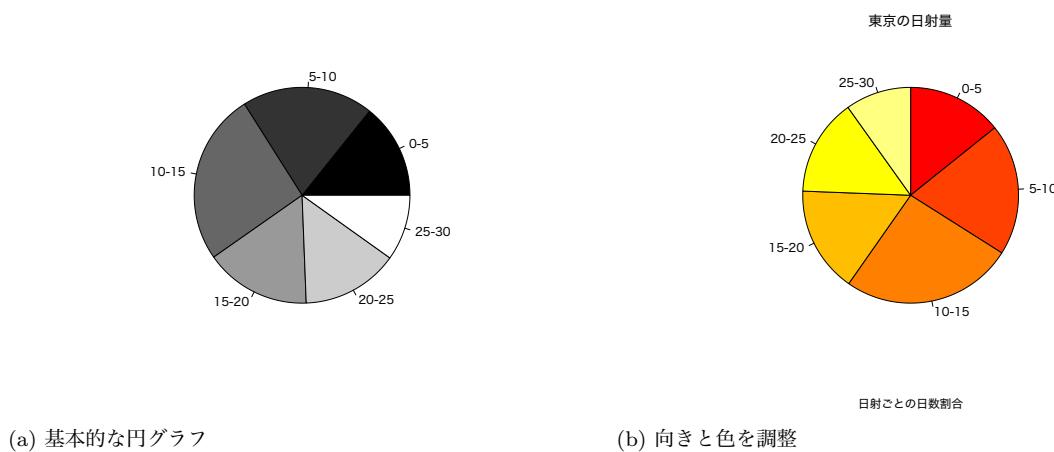


図 4.7: 関数 pie() の例.

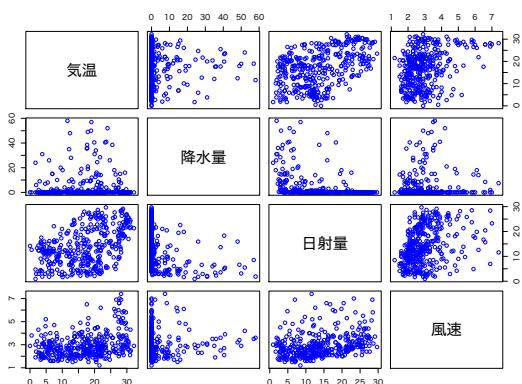
Rscript: graph-pie.r

図 4.7 参照

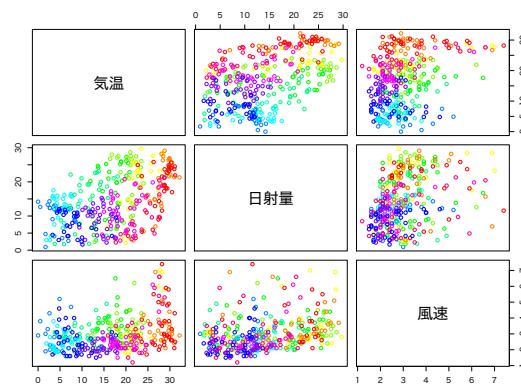
```
> ### 関数 pie による円グラフの作図
> par(family="HiraginoSans-W4")
> myData <- subset(read.csv("data/tokyo_weather.csv",
+                           fileEncoding="utf8"),
+                           select=c(月, 気温, 降水量, 日射量, 風速))
> z <- hist(myData$日射量, breaks=5, # 5つ程度に分類
+            plot=FALSE) # ヒストグラムのBINの情報のみ取得
> x <- z$count # 各BINの頻度
> y <- z$breaks # BINの境界 (BINの数より1つ多い)
> names(x) <- paste(y[-length(y)], y[-1], sep="-")
> ### 基本的な円グラフ
> pie(x, col=gray(seq(0,1,length=length(x))))
> ### 向きと色を調整
> pie(x,
+       clockwise=TRUE, # 時計まわりで12時から描画
+       col=heat.colors(length(x)),
+       main="東京の日射量", sub="日射ごとの日数割合")
```

4.6 散布図行列

多次元データの変数間の関係を概観するために、2つの変数間の散布図を複数行列状に並べた図を用いることがある。これは関数 `pairs()` によって作成することができる（関数 `plot()` でも同じことができる）。



(a) 基本的な散布図



(b) 表示する項目を指定

図 4.8: 関数 `pairs()` の例。

Rscript: graph-pairs.r

図 4.8 参照

```
> ### 関数 pairs による散布図の作図
> par(family="HiraginoSans-W4")
> myData <- subset(read.csv("data/tokyo_weather.csv",
+                           fileEncoding="utf8"),
+                           select=c(月, 気温, 降水量, 日射量, 風速))
> ### 基本的な散布図
> pairs(myData[,-1], col="blue")
> ## plot(myData[,-1], col="blue") でも同じ図が描ける
> ### 表示する項目を指定
> myColor <- rainbow(12)[c(8:1,12:9)] # 月別の色を用意
```

```
> pairs(~ 気温 + 日射量 + 風速, data=myData,
+       col=myColor[myData$月]) # 月毎に異なる色で表示
```

4.7 3次元のグラフ

3次元のグラフを2次元に射影した俯瞰図は、関数 `persp()` を用いて描くことができる。視線の方向はオプション `theta` と `phi` で極座標を指定することによって制御することができる。パッケージ `scatterplot3d` には、3次元の散布図を書くための関数 `scatterplot3d()` が用意されている。

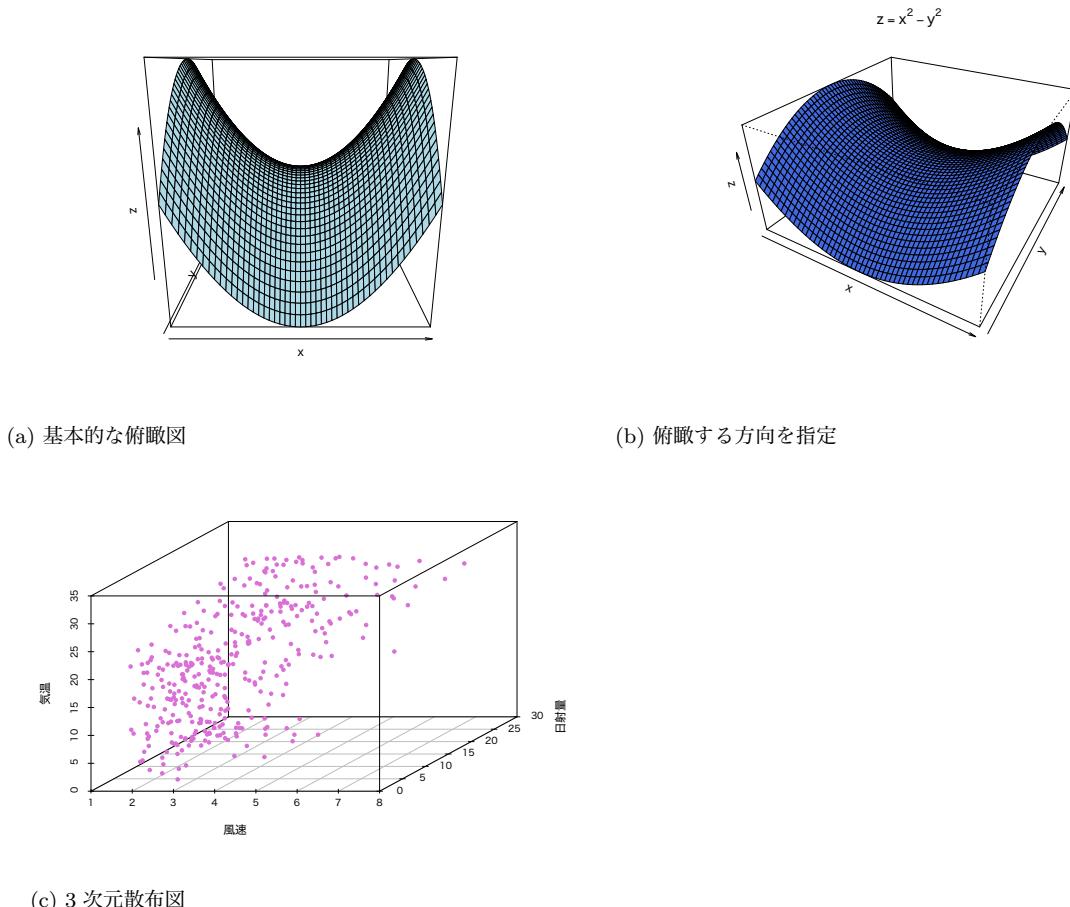


図 4.9: 3次元グラフの例。

Rscript: [graph-plot3d.r](#)

図 4.9 参照

```
> ### 関数 persp による 2変数関数の俯瞰図
> f <- function(x,y) x^2 - y^2
> x <- seq(-3, 3, length=51) # x 座標の定義域の分割
> y <- seq(-3, 3, length=51) # y 座標の定義域の分割
> z <- outer(x, y, f) # z 座標の計算
> ### 基本的な俯瞰図
> persp(x, y, z, col="lightblue")
> ### 俯瞰する向きを指定
> persp(x, y, z,
+        theta=30, phi=30, expand=0.5, # 視点を設定
```

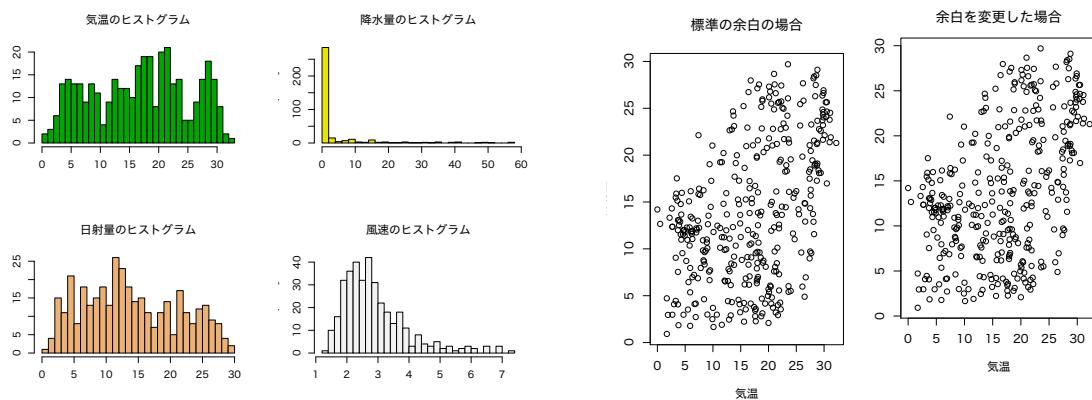
```

+           col="royalblue", main=expression(z==x^2-y^2))
> ### 3次元散布図(パッケージ scatterplot3d を利用)
> ## install.packages("scatterplot3d") # 初めて使う場合
> library(scatterplot3d) # パッケージのロード
> par(family="HiraginoSans-W4")
> myData <- subset(read.csv("data/tokyo_weather.csv",
+                           fileEncoding="utf8"),
+                           select=c(風速, 日射量, 気温))
> scatterplot3d(myData, pch=20, color="orchid")
> ## pch の指定については help(points) を参照

```

4.8 プロット環境の設定

プロットの際の線の種類や色、点の形等のデフォルト値は関数 `par()` で設定できる。設定可能なグラフィックスパラメータは `help(par)` で確認できる。特に以下の例のように関数 `par()` によってプロット環境の設定(複数図の配置、余白の設定など)ができる。



(a) 複数の図の配置例

(b) 余白の指定の例

図 4.10: 環境の設定の例。

Rscript: `graph-par.r`

図 4.10 参照

```

> ### 複数図の配置
> ## 気候データの各変数のヒストグラムを1つの画面に配置
> par(family="HiraginoSans-W4")
> myData <- subset(read.csv("data/tokyo_weather.csv",
+                           fileEncoding="utf8"),
+                           select=c(気温, 降水量, 日射量, 風速))
> op <- par(mfrow=c(2,2)) # 画面を2x2に分割、行方向に配置
> ## par(mfcol=c(2,2)) で列方向に配置できる
> myColor <- terrain.colors(4) # 色を用意
> myName <- colnames(myData) # ヒストグラムの変数名
> ## 第1変数のヒストグラムの作成
> hist(myData[,myName[1]], col=myColor[1],
+       breaks=25, xlab="",
+       main=paste0(myName[1], "のヒストグラム"))
> ## 第2変数のヒストグラムの作成
> hist(myData[,myName[2]], col=myColor[2],
+       breaks=25, xlab="",
+       main=paste0(myName[2], "のヒストグラム"))

```

```

> ## 残りは for 文で作成
> for(i in 3:4){
+   hist(myData[,myName[i]], col=myColor[i],
+         breaks=25, xlab="", 
+         main=paste0(myName[i], "のヒストグラム"))
+ }
> par(op) # 設定解除 (古い設定 <- par(新しい設定))
> ### 余白の設定
> op1 <- par(mfrow=c(1,2))
> plot(myData[,c(1,3)], main="標準の余白の場合")
> op2 <- par(mar=c(7,1,3,5))
> ## 下・左・上・右の順で余白を設定
> ## デフォルトは par(mar=c(5,4,4,2)+0.1)
> plot(myData[,c(1,3)], main="余白を変更した場合")
> par(op2); par(op1) # 設定解除

```

演習 4.1. 前章で整理したデータフレームを描画してみよう.

1. 関数 `data()` で調べた適当なデータフレームを描画しなさい.
2. 上記のデータフレームを集計し、その結果を描画しなさい.

4.9 補遺

4.9.1 参考文献

この章に関連する参考書としては以下を挙げておく.

- [1] 金明哲. *Rによるデータサイエンス(第2版)*. 東京: 森北出版, 2017.
- [2] 奥村晴彦. *Rで楽しむ統計*. 東京: 共立出版, 2016.

4.9.2 パッケージ `ggplot2` の利用

データの可視化は、データ解析において基本的かつ有効な方法であるのみならず、分析結果を他の人々に説明する際の資料としても必須のものである。そのため R のグラフィック機能を拡張するためのパッケージも多数開発されている。その中でも、近年利用が広まっているものにパッケージ `ggplot2` がある。`ggplot2` は、統一的な文法で系統的に美しいグラフを描くことを目的として開発されているパッケージである。基本設計は確定しているが、細かい部分は現在も頻繁に開発が進められている。用意されている関数の細かな情報については、

<https://docs.ggplot2.org/>

に詳しい例題とともにまとめられている。また、良く使われる関数については、簡潔に纏めた 2 頁のシート

<https://rstudio.com/wp-content/uploads/2015/03/ggplot2-cheatsheet.pdf>

が用意されているので、興味に応じて参照してほしい。

4.9.3 基本的な文法

`ggplot2` では、どのデータを対象とし、どの変数を座標とし、どのような図を描くのかを順に指定するといった思想で、文法が設計されている。

まず、どのデータを対象とするかを指定すると同時に、2次元のグラフの x 軸(横軸)と y 軸(縦軸)に何を用いるかを指定する必要がある。いくつか方法は用意されているが、関数 `ggplot2::ggplot()` を用いるのが標準である。関数 `ggplot2::ggplot()` を用いる場合には、データフレームを渡すとともに、関数 `aes()` を用いて x 座標と y 座標に対応する変数を指定することができる。関数 `ggplot2::ggplot()` だけでは何も描画はされないが、以降の描画ではこれらが既定値(指定しなくとも自動的に用いられる値)として使われることになる。この後、どのようなグラフを描くかは描画の内容を指示する関数を付与(+で加えていく)して指定するのが基本的な文法となる。

4.9.4 散布図

最も基本的な図は2次元の散布図(scatterplot)であるが、これには指定したデータ点を描画する関数 `ggplot2::geom_point()` を用いる。点の大きさや色も座標の一種と考えることができ、これらを使うと2次元以上の情報を視覚化することもできる。また、図のタイトルなどは関数 `labs()` を用いて指定することができる。

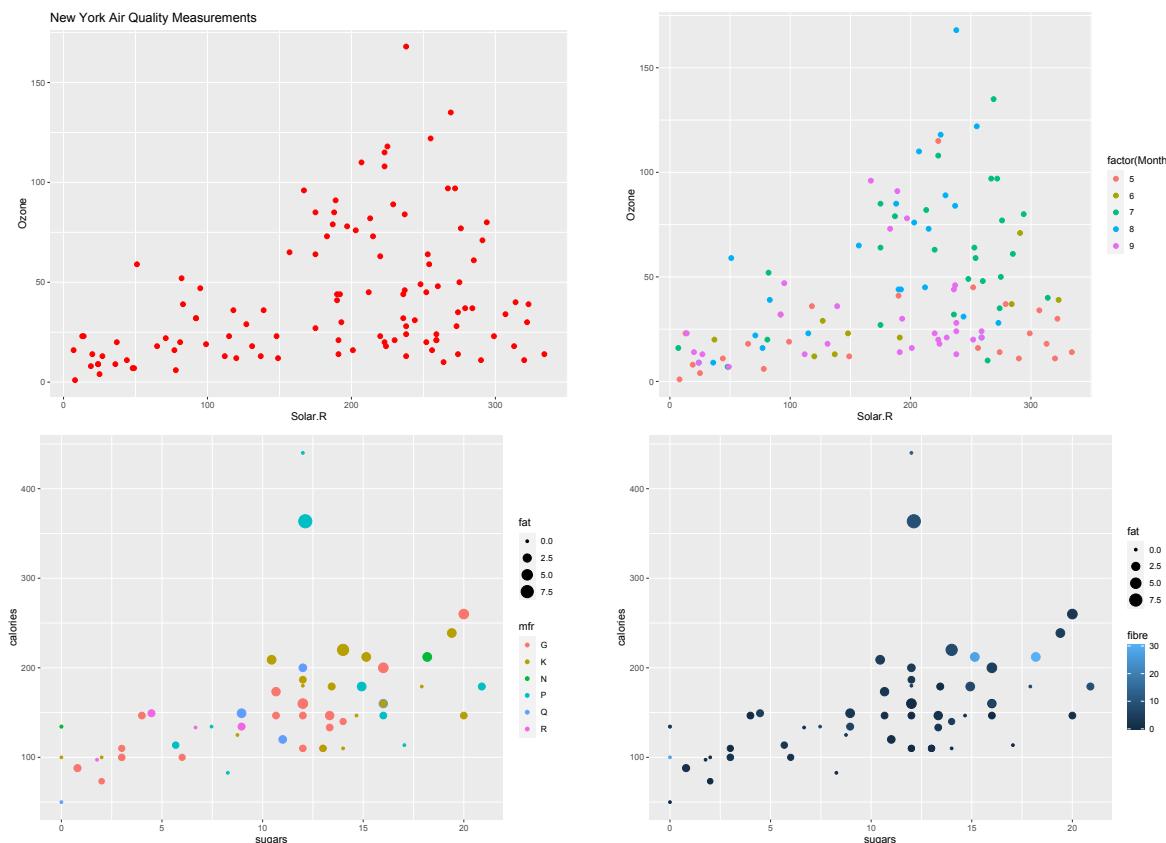


図 4.11: 関数 `geom_point()`.

図 4.11 参照

```
> ### datasets::airquality を用いた描画例
> library(MASS) # パッケージの読み込み
> library(tidyverse) # 読み込む順番にも注意が必要
> ## 日射量とオゾン量の関係を見るために散布図を描く
> ggplot(airquality, aes(Solar.R, Ozone)) +
+   geom_point(colour="red", size=2, # 色とサイズを指定
+              na.rm=TRUE) +
+   labs(title="New York Air Quality Measurements")
> ggplot(airquality, aes(Solar.R, Ozone)) +
+   geom_point(aes(colour=factor(Month)), # 月毎に色分け
+              size=2, na.rm=TRUE)
> ### MASS::UScereal を用いた描画例
> ## 会社別に糖分、カロリー、脂肪分の関係性を見る
> ggplot(UScereal, aes(sugars, calories)) +
+   geom_point(aes(size=fat, # サイズと脂肪分を対応付け
+                  colour=mfr)) # 色と会社を対応付け
> ## 糖分、カロリー、脂肪分、繊維質の関係性を見る
> ggplot(UScereal, aes(sugars, calories)) +
+   geom_point(aes(size=fat,
+                  colour=fibre)) # 色と繊維質を対応付け
```

Rscript: [ggplot-point.r](#)

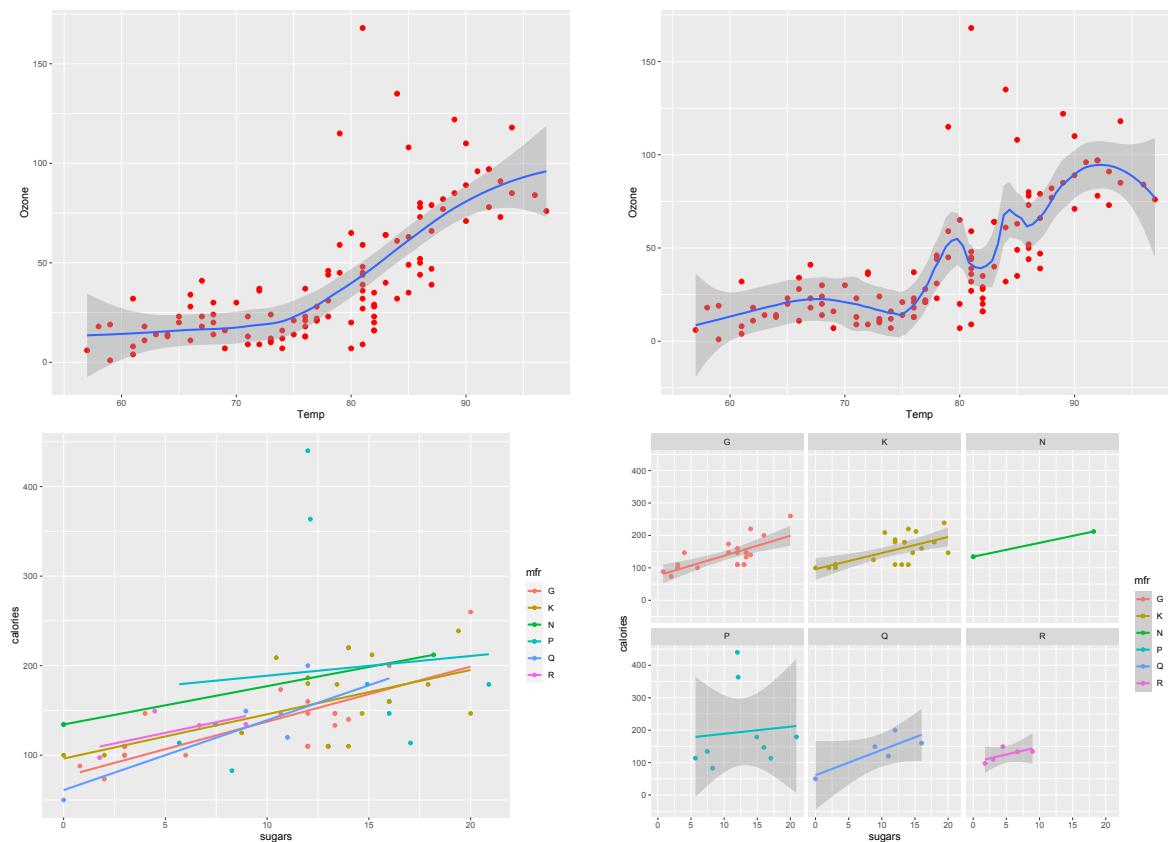
4.9.5 曲線あてはめ

データの x 座標と y 座標との間にある関係を視覚化するには、曲線あてはめを行うのが簡便である。データ点に適当な方法であてはめた曲線は関数 `ggplot2::geom_smooth()` によって描くことができる。既定値では `loess` 法を用いた曲線と標準誤差から求められる信頼区間が描かれる。オプション `method` に”`lm`”を指定することによって直線あてはめ(線形回帰)を行うこともできる。

図 4.12 参照

```
> ### datasets::airquality を用いた描画例
> ## 溫度とオゾン量の関係を回帰曲線を描く
> ggplot(airquality, aes(Temp, Ozone)) +
+   geom_point(colour="red", size=2, na.rm=TRUE) +
+   geom_smooth(na.rm=TRUE)
> ## 曲線の滑らかさを変える
> ggplot(airquality, aes(Temp, Ozone)) +
+   geom_point(colour="red", size=2, na.rm=TRUE) +
+   geom_smooth(span=0.3, na.rm=TRUE) # 幅の狭い平滑化
> ### MASS::UScereal を用いた描画例
> ## 会社毎の糖分とカロリーの回帰直線を見る
> ggplot(UScereal, aes(sugars, calories, colour=mfr)) +
+   geom_point() +
+   geom_smooth(method="lm", se=FALSE) # 信頼区間無し
> ## 信頼区間を付けて別々に表示する
> ggplot(UScereal, aes(sugars, calories, colour=mfr)) +
+   geom_point() +
+   geom_smooth(method="lm") +
+   facet_wrap(~ mfr) # 会社ごとに別のグラフを作成
```

Rscript: [ggplot-smooth.r](#)

図 4.12: 関数 `geom_smooth()`.

4.9.6 ヒストグラム

1次元データの分布の概形を捉えるにはヒストグラム (histogram) が重要であるが、関数 `ggplot2::geom_histogram()` を用いることによって描画することができる。また、棒状のグラフではなく折れ線で描くこともでき、これには関数 `ggplot2::geom_freqpoly()` を用いる。なお、 y 軸の値は自動的に計算されるので特に指定する必要はないが、複数の値を比較する場合には密度に正規化して表示した方が良いこともある。内部で計算された密度の値 (`density`) を y 軸に指定するには `..density..` を指定する。

Rscript: `ggplot-hist.r`

図 4.13 参照

```
> ### datasets::airquality を用いた描画例
> ## オゾン量のヒストグラムを描く
> ggplot(airquality, aes(Ozone)) +
+   geom_histogram(bins=25, # ビンの数を指定
+   na.rm=TRUE)
> ## オゾン量のヒストグラムを折れ線で描く
> ggplot(airquality, aes(Ozone)) +
+   geom_freqpoly(binwidth=5, # ビンの幅を指定
+   colour="blue", na.rm=TRUE)
> ### MASS::UScereal を用いた描画例
> ## 会社毎のカロリーを詰み上げてヒストグラムを描く
> ggplot(UScereal, aes(calories)) +
+   geom_histogram(binwidth=30, aes(fill=mfr))
> ## 会社別の頻度(..density..)を折れ線で描く
> ggplot(UScereal,
+   aes(calories, ..density.., colour=mfr)) +
```

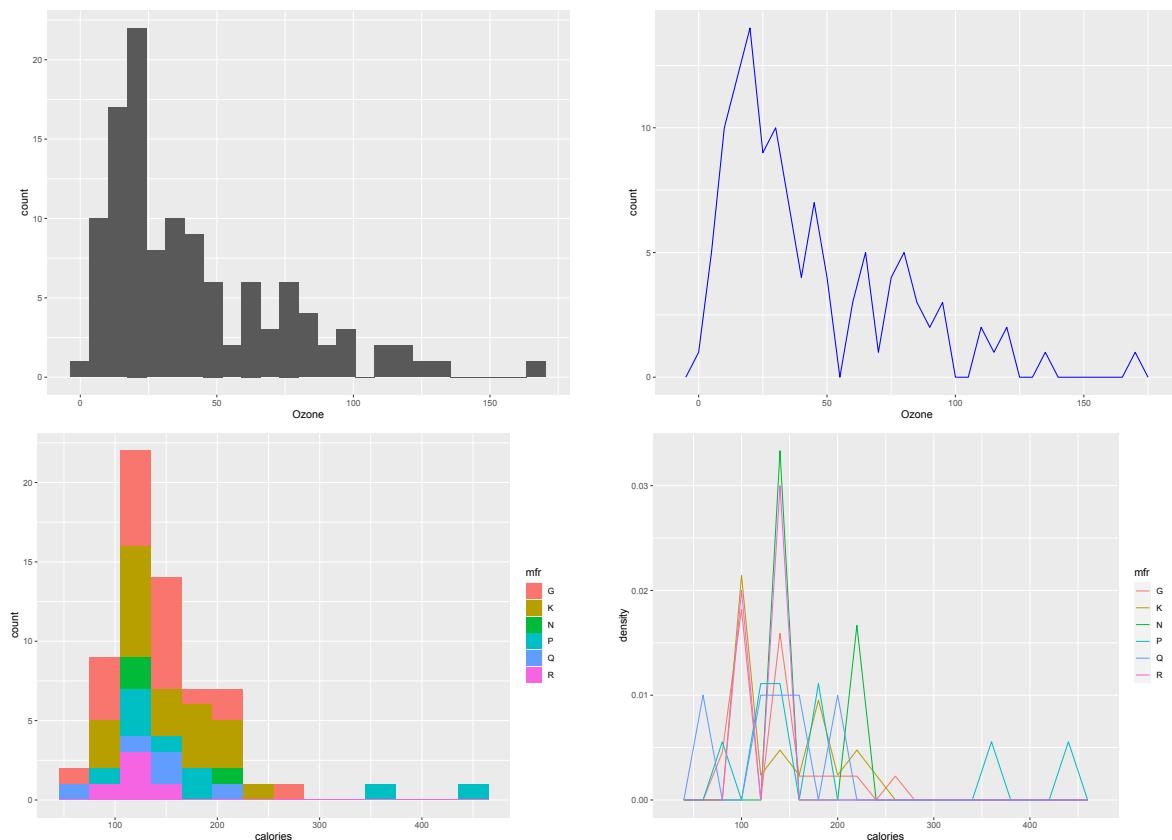


図 4.13: 関数 geom_histogram()。

```
+ geom_freqpoly(binwidth=20)
```

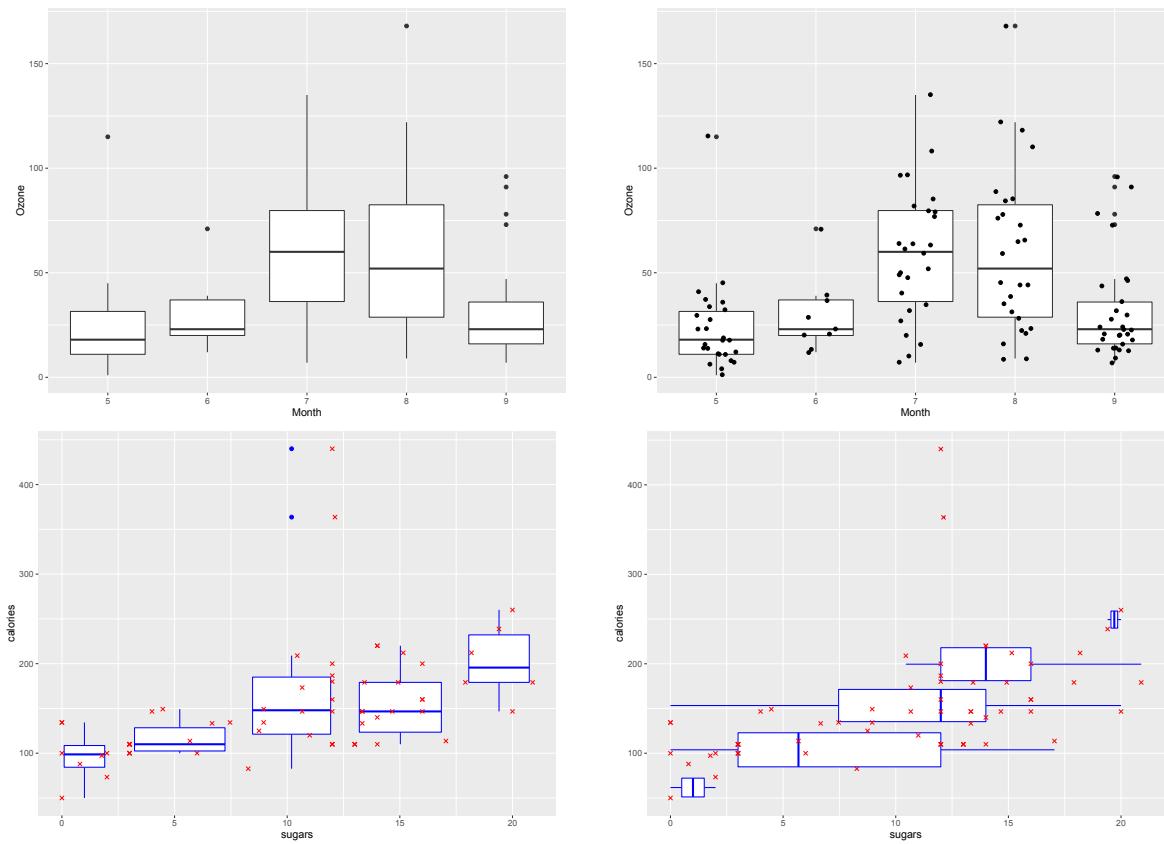
4.9.7 箱ひげ図

ヒストグラムより簡便に分布の様子を視覚化する方法として箱ひげ図 (boxplot) がある。これは関数 `ggplot2::geom_boxplot()` を用いて描くことができる。

図 4.14 参照

```
> ### datasets::airquality を用いた描画例
> ## 月別のオゾン量の箱ひげ図を描く
> ggplot(airquality, aes(factor(Month), Ozone)) +
+   geom_boxplot(na.rm=TRUE) +
+   labs("x"="Month") # x 軸のラベルを変更
> ## データを付随させる
> ggplot(airquality, aes(factor(Month), Ozone)) +
+   geom_boxplot(na.rm=TRUE) +
+   geom_jitter(width=0.2, na.rm=TRUE) + # jitter は
+   labs("x"="Month") # データをずらして重ねない操作
> ### MASS::UScereal を用いた描画例
> ## 糖分とカロリーの関係性を見る
> ggplot(UScereal, aes(sugars, calories)) +
+   geom_boxplot(aes(group=cut_width(sugars,5)),
+               colour="blue") +
+   geom_point(colour="red", shape=4)
> ## 軸を入れ換えて箱ひげ図を描く
> ggplot(UScereal, aes(calories, sugars)) + # xy を交換
```

Rscript: `ggplot-boxplot.r`

図 4.14: 関数 `geom_boxplot()`.

```
+ geom_boxplot(aes(group=cut_width(calories,50)),
+               colour="blue") +
+ geom_point(colour="red", shape=4) +
+ coord_flip() # 軸の入れ換えて xy の配置を戻す
```

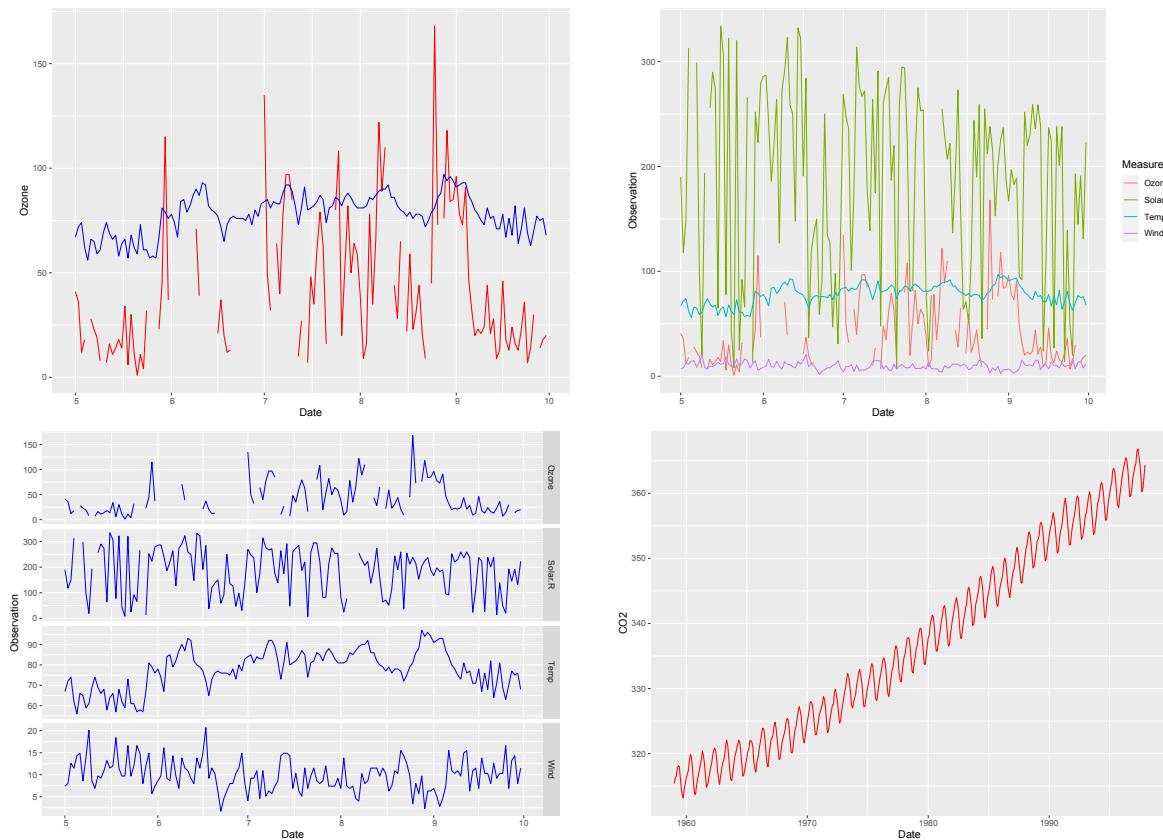
4.9.8 折れ線グラフ

折れ線グラフは時系列など x 軸の増加に伴なう y 軸の変動を視覚化する場合に用いられる。点列を結ぶ線を描くには関数 `ggplot2::geom_line()` を用いる。なお、時系列を図示する際に、データフレームに時間の情報が不足している場合には自分で時間に関する変数を整理し直さなくてはならないので、注意が必要である。

Rscript: `ggplot-line.r`

図 4.15 参照

```
> ### datasets::airquality を用いた描画例
> ## オゾン量を時系列としてグラフを描く
> myData <- airquality %>%
+   mutate(Date=as.Date( # 時間の情報を整理して列を作成
+     paste(Month, Day, "73", sep="/"),
+     "%m/%d/%y")) %>%
+   select(Date, Ozone:Temp)
> ggplot(myData, aes(x=Date)) +
+   geom_line(aes(y=Ozone), colour="red") +
+   geom_line(aes(y=Temp), colour="blue") # 重ね描き
> ## 複数の系列を描画
> library(tidyverse)
```

図 4.15: 関数 `geom_line()`.

```
> myData <- airquality %>%
+   mutate(Date=as.Date(
+     paste(Month, Day, "73", sep="/"),
+     "%m/%d/%y")) %>%
+   select(Ozone:Temp, Date) %>%
+   gather(Measure, Observation, Ozone:Temp)
> ## 
> ggplot(myData, aes(Date, Observation)) +
+   geom_line(aes(colour=Measure)) # データ毎に色を変更
> ## 別の facet に表示
> ggplot(myData, aes(Date, Observation)) +
+   geom_line(colour="blue") +
+   ## 計測データ毎に別のグラフ (facet) で表示
+   facet_grid(Measure ~ ., scales="free_y") # y を調整
> ### datasets::co2 を用いた描画例
> ## ts class のデータを変換して描画
> myData <- data.frame(CO2=as.vector(co2),
+                       Date=as.vector(time(co2)))
> ggplot(myData, aes(Date,CO2)) +
+   geom_line(colour="red")
```

4.9.9 棒グラフ

棒グラフは集計されたデータの比較を視覚的に行う場合に重要であるが、これは関数 `ggplot2::geom_bar()` を用いて描画することができる。データフレームの集計は、前章で用いたパッケージ `dplyr`などを用いて行うことになる。

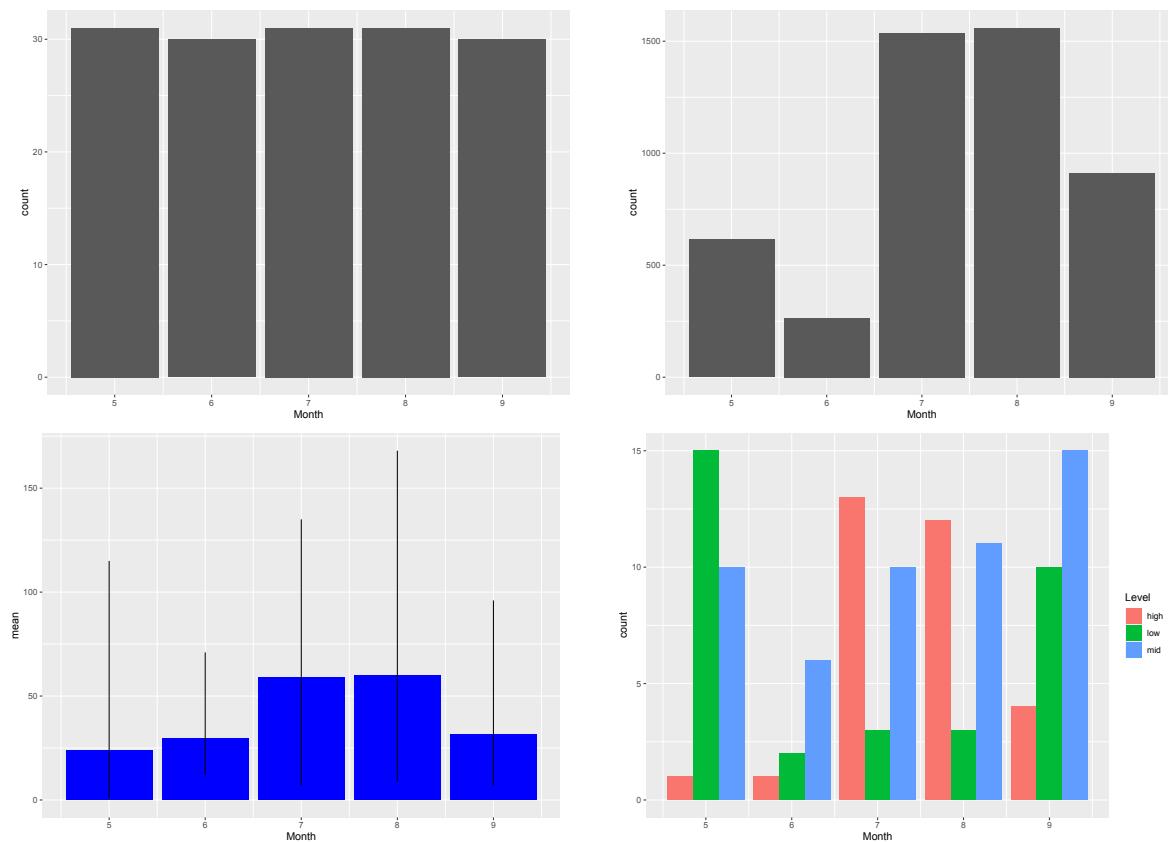


図 4.16: 関数 geom_bar()。

Rscript: [ggplot-bar.r](#)**図 4.16 参照**

```
> ### datasets::airquality を用いた描画例
> ## 月別の計測日数を棒グラフとして表示
> ggplot(airquality, aes(Month)) +
+   geom_bar()
> ## 月別のオゾン量の合計を表示
> ggplot(airquality, aes(Month)) +
+   geom_bar(aes(weight=Ozone))
> ## 月別にオゾン量を集計
> myData <- airquality %>%
+   group_by(Month) %>%
+   summarize(
+     mean      = mean(Ozone, na.rm=TRUE),
+     min       = min(Ozone, na.rm=TRUE),
+     max       = max(Ozone, na.rm=TRUE))
> ## 月別のオゾン量の平均を表示
> ggplot(myData, aes(Month, mean)) +
+   geom_bar(stat="identity", fill="blue") +
+   geom_linerange(aes(ymin=min,ymax=max)) # 値域も表示
> ## オゾン量の高低のラベルを付加
> myData <- airquality %>%
+   mutate(Level=ifelse(Ozone>60, "high",
+                      ifelse(Ozone<20, "low", "mid"))) %>%
+   filter(!is.na(Level)) # 計算できなかったものを除く
> ## 月別のオゾン量の高低の割合を表示
> ggplot(myData, aes(Month)) +
+   geom_bar(aes(fill = Level), position = "dodge")
```

モンテカルロ法

現実の世界には確率的な取り扱いが必要な事象があり、それらを確率的現象という。現実のデータに含まれる不確定性は確率的現象によって生まれる。確率的現象にはさまざまなものがあり、中には我々の直感と大きく異なる現象も含まれる。確率的現象は、問題を抽象化・単純化して理論的な解析を詳しく行うことが可能な場合もあるが、理論的に解析を行うことが難しい現象も数多く存在する。そうした複雑な問題に対して、計算機上の擬似乱数を利用して数値的に現象を再現し、その性質を調べる方法がある。それは**モンテカルロ法** (Monte-Carlo method) あるいは**確率シミュレーション** (stochastic simulation) と呼ばれる。計算機上では繰り返しシミュレーションを行うことができるるので、原因となる要素が変化すると結果にどのような影響を及ぼすかを詳細に調べることができる。この章では簡単な例を取り上げながらモンテカルロ法の考え方について紹介する。

5.1 亂数

亂数とはランダムに生成された数列のことである。もちろんコンピューターでは完全にランダムに数字を発生させることは不可能なため、それらの乱数は厳密には**擬似乱数**と呼ばれる。¹特に数値シミュレーションを行う上では、それが再現可能であることが要請されるため、発生される乱数も再現可能である必要がある。**R**ではこれを実行するために、乱数の初期値を指定するための関数 `set.seed()` が用意されている(同一の初期値から生成される乱数は同一のものとなる)。

ここでは基本的な乱数として、ランダムサンプリング、二項乱数および一様乱数を考える。ランダムサンプリングは、その名の通り「与えられた集合の要素をランダムに抽出することで発生」する乱数のことである。二項乱数は、「確率 p で表ができるコインを n 回投げた際の表が出る回数」に対応する乱数である。従って p と n によって乱数の発生の仕方が変わるため、それを明示する場合は「確率 p に対する次数 n の二項乱数」と言う。一様乱数は、「ある決まった区間 (a, b) ($a < b$) に含まれる数字からランダムに発生」する乱数のことである¹。従って区間 (a, b) によって乱数の発生の仕方が変わるために、それを明示する場合は「区間 (a, b) 上の一様乱数」と言う。

ランダムサンプリングは関数 `sample()` で実行できる。二項乱数および一様乱数はそれぞれ関数 `rbinom()` および `runif()` を用いて発生させる。

```
> ### 関数 sample の使い方
> x <- 1:10 # サンプリング対象の集合をベクトルとして定義
> set.seed(123) # 亂数のシード値(任意に決めてよい)を指定
> sample(x, # x から
```

¹ R では擬似乱数を発生させるための方法として “Mersenne-Twister” がデフォルトでは用いられている。`help(Random)` 参照。

Rscript: `mc-sample.r`

¹ (a, b) は a より大きく b より小さい実数全体からなる集合を表す。

```

+      5) # 5つの要素を重複なしでランダムに抽出
[1] 3 10 2 8 6

> sample(x, # xの要素のランダムな並べ替えとなる
+         length(x)) # (要素数と同じ数抽出)
[1] 5 4 6 8 1 2 3 7 9 10

> sample(x, 5,
+         replace=TRUE) # 重複ありでランダムに抽出
[1] 9 9 9 3 8

> sample(1:6, 10, replace=TRUE) # サイコロを 10 回振る
[1] 2 2 1 6 3 4 6 1 3 5

> sample(1:6, 10, prob=6:1, # 出る目の確率(比率)に偏り
+         replace=TRUE)
[1] 1 1 2 2 2 1 1 1 2 1

> ### 関数 rbinom の使い方
> rbinom(10, # 二項乱数を 10 個発生
+         size=4, # 次数 4 の二項乱数
+         prob=0.5) # 1 となる確率が 0.5
[1] 3 0 2 3 1 2 1 1 3 3

> rbinom(20, # 個数を 20 に変更
+         size=4,
+         prob=0.2) # 確率を 0.2 に変更
[1] 0 1 0 0 0 1 1 1 1 1 1 1 1 1 0 1 0 0 1 0

> ### 関数 runif の使い方
> runif(5, # 一様乱数を 5 個発生
+         min=-1, max=2) # 区間 (-1, 2) 上
[1] -0.6665937 -0.2691416 1.0041668 0.2529403
[5] 1.3645875

> runif(5) # 指定しない場合は区間 (0, 1) が既定値
[1] 0.1028646 0.4348927 0.9849570 0.8930511 0.8864691

> ### 関数 set.seed について
> set.seed(1) # 亂数の初期値を seed=1 で指定
> runif(5)
[1] 0.2655087 0.3721239 0.5728534 0.9082078 0.2016819

> set.seed(2) # 亂数の初期値を seed=2 で指定
> runif(5) # seed=1 の場合と異なる結果
[1] 0.1848823 0.7023740 0.5733263 0.1680519 0.9438393

> set.seed(1) # 亂数の初期値を seed=1 で指定
> runif(5) # 初めの seed=1 の場合と同じ結果
[1] 0.2655087 0.3721239 0.5728534 0.9082078 0.2016819

```

R には他にも様々な種類の確率分布に従う乱数が実装されて

いる。

5.2 数値シミュレーション

以下では具体的な例題を用いて確率的なシミュレーションを説明する。

5.2.1 コイン投げの賭け

まず初めに次の簡単な問題を考えてみよう。

問題 5.1. A と B の二人で交互にコインを投げる。最初に表が出た方を勝ちとするとき、A と B それぞれの勝率はいくつとなるか？

コインを投げる試行は Bernoulli 分布(サイズ 1 の 2 項分布)なので、乱数生成には関数 `rbinom()` を利用することができる。

別の方法としては関数 `runif()` を利用して生成した乱数が $1/2$ 以上であるかどうかで模擬することもできる。

シミュレーションの一例を以下に示す。

```
> ### コイン投げの賭け
> ## コイン投げの試行
> ## いろいろな書き方があるが、まずはベタに書いてみる
> myTrial <- function() { # 名前は何でも良い
+   while(TRUE){ # 永久に回るループ
+     if(rbinom(1,size=1,prob=0.5)==1){
+       return("A") # A が表を出して終了
+     }
+     if(rbinom(1,size=1,prob=0.5)==1){
+       return("B") # B が表を出して終了
+     }
+     ## どちらも裏ならもう一度ループ
+   }
+ }
> ## 試行を行ってみる
> myTrial()

[1] "A"
> myTrial()
[1] "A"
> myTrial()
[1] "A"

> ## Monte-Carlo simulation
> ## set.seed(8888) # 実験を再現する場合はシードを指定
> mc <- 10000 # 回数を設定
> myData <- replicate(mc,myTrial())
> ## 簡単な集計
> table(myData) # 頻度

myData
  A    B
6705 3295

> table(myData)/mc # 確率(推定値)

myData
  A    B
0.6705 0.3295
```

Rscript: `mc-coin.r`

5.2.2 Buffon の針

次に 18 世紀の学者 Buffon による有名な問題を考えてみよう。

問題 5.2. 2 次元平面上に等間隔 d で平行線が引いてある。長さ l の針 ($l < d$ とする) をこの平面上にランダム(でたらめ)に落としたとき、平行線と交わる確率はいくつか？

針の中心位置と一番近い平行線を原点とし、水平方向の座標を x とする。問題の繰り返し構造に注意すれば、 x は区間 $[-d/2, d/2]$ で一様に分布する(どの点が得られるかは無作為)と考えられる。また針に向きがあるとして、針が図の水平方向となす角度を θ とする。 θ も同様に区間 $[0, 2\pi]$ で一様に分布すると考えられる。この試行の見本点は (x, θ) で表され、その見本空間は

$$\Omega = [-d/2, d/2] \times [0, 2\pi] \subset \mathbb{R}^2$$

となる。また、針が平行線と交わる条件は、針の両端の座標の符号が異なること

$$\left(x + \frac{l}{2} \cos \theta\right) \left(x - \frac{l}{2} \cos \theta\right) < 0$$

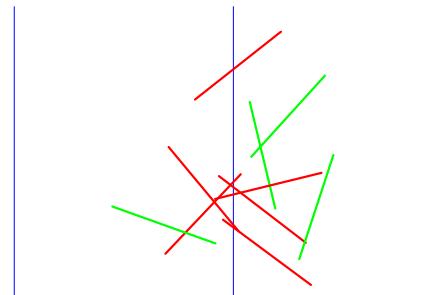
で表される。

上記の条件を満たす (x, θ) の領域を考えると

$$P(\text{針が平行線と交わる}) = \frac{4l}{2\pi d} = \frac{2l}{\pi d}$$

となる。

図 5.1: 針と平行線の関係を図示した例。



Rscript: `mc-buffon.r`

図 5.1 参照

```
> ### Buffon の針
> ## 針を投げる試行
> myTrial <- function(d,l,verbose=FALSE){ # d と l を指定
+   x <- runif(1,min=-d/2,max=d/2) # 位置
+   theta <- runif(1,min=0,max=2*pi) # 角度
+   cross <- FALSE # 交わったかどうかを示す変数
+   if((x+l*cos(theta)/2)*(x-l*cos(theta)/2)<0){
+     cross <- TRUE # 交わった場合に書き換え
+   }
+   if(verbose==TRUE){ # 位置と角度も返す
```

```

+
+         return(c(x=x, theta=theta,
+                     cross=as.numeric(cross)))
+     } else { # 交わったかどうかだけ返す
+         return(cross)
+     }
+ }
> ## 試行を行ってみる
> d <- 10
> l <- 5
> myTrial(d,l,verbose=TRUE)

      x      theta      cross
-2.592678  1.323649  0.000000

> myTrial(d,l,verbose=TRUE)

      x      theta      cross
3.868208  2.642052  0.000000

> ## 絵にしてみる
> plot(c(0,0),type="n", # 空のキャンバスを作る
+       xlim=c(-d,d),asp=1,ann=FALSE,axes=FALSE)
> abline(v=c(-10,0,10),col="blue") # 線を引く
> for (i in 1:10) {
+   obs <- myTrial(d,l,verbose=TRUE)
+   x <- obs["x"]
+   theta <- obs["theta"]
+   y <- runif(1,min=-d/2,max=d/2) # y座標はランダム
+   x1 <- x-1/2*cos(theta)
+   x2 <- x+1/2*cos(theta)
+   y1 <- y-1/2*sin(theta)
+   y2 <- y+1/2*sin(theta)
+   lines(c(x1,x2),c(y1,y2),
+         col=ifelse(x1*x2<0,"red","green"),
+         lty="solid", lwd=3)
+ }
> ## Monte-Carlo simulation
> ## set.seed(8888) # 実験を再現したい場合はシードを指定
> mc <- 10000 # 回数を設定
> myData <- replicate(mc,myTrial(d,l))
> ## 簡単な集計
> table(myData) # 頻度 (TRUE が針の交わった回数)

myData
FALSE  TRUE
 6832  3168

> table(myData)/mc # 確率 (推定値)

myData
 FALSE  TRUE
 0.6832 0.3168

> print((2*l)/(pi*d)) # 針の交わる確率 (理論値)

[1] 0.3183099

```

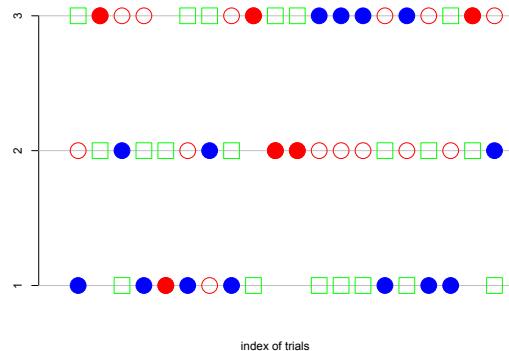
5.2.3 Monty Hall 問題

次の問題はアメリカの雑誌で、その解をめぐって大議論に発展した問題である。

問題 5.3. プレーヤーの前に閉まった3つのドアがある。1つのドアの後ろには景品の新車が、2つのドアの後ろにははずれを意味するヤギがいる。プレーヤーは新車のドアを当てることができると景品として新車がもらえる。プレーヤーが1つのドアを選択した後、司会のモンティが残りのドアのうちヤギがいるドアを開けてヤギを見せる。(プレーヤーがどのドアを選んでも、モンティはヤギのいるドアを開けられることに注意せよ。) ここでプレーヤーは、最初に選んだドアを、開けられていない残ったドアに変更してもよいと言われる。プレーヤーはドアを変更すべきだろうか？

実際にシミュレーションを行うと以下のようになり、最初の選択と変更した場合で景品を貰える確率が異なることがわかる。

図 5.2: シミュレーション例: 赤○は最初に選択したドアの位置、緑四角は開けられたドアの位置、赤塗り潰しは最初の選択のままが正解の場合、青塗り潰しはドアを変えるのが正解の場合。



Rscript: [mc-montyhole.r](#)

図 5.2 参照

```
> ### Monty Hole 問題
> ## クイズに答える試行
> myTrial <- function(verbose=FALSE){
+   prize <- sample(1:3,size=1) # 賞品の置かれた扉
+   choice <- sample(1:3,size=1) # 最初の選択
+   if(prize==choice) { # 変えないのが正解
+     win <- "stay"
+     door <- sample(setdiff(1:3,prize),size=1)
+   } else { # 変えるのが正解
+     win <- "change"
+     door <- setdiff(1:3,union(prize,choice))
+   }
+   if(verbose==TRUE){ # 賞品, 選択, 正しい扉を返す
+     return(c(prize=prize,choice=choice,door=door))
+   } else { # 勝ち負けの条件を返す
+     return(win)
+   }
+ }
> ## 試行を行ってみる
> myTrial()
[1] "stay"

> myTrial(verbose=TRUE)

prize choice   door
1         2       3

> ## 絵にしてみる
> mc <- 20
```

```

> plot(c(0,0), type="n", # 空のキャンバスを作る
+       xlim=c(0,mc), ylim=c(1,3), ann=FALSE, axes=FALSE)
> title(xlab="index of trials")
> axis(2, at=1:3, labels=1:3)
> abline(h=1:3, col="grey") # 線を引く
> for (i in 1:mc) {
+   obs <- myTrial(verbose=TRUE)
+   prize <- obs["prize"]
+   choice <- obs["choice"]
+   door <- obs["door"]
+   points(i,door,pch=0,cex=3,col="green")
+   points(i,prize,pch=1,cex=3,col="red")
+   points(i,choice,pch=19,cex=3, # 正解を色で表示
+          col=ifelse(prize==choice, "red", "blue"))
+ }
> ## Monte-Carlo simulation
> ## set.seed(8888) # 実験を再現したい場合はシードを指定
> mc <- 10000 # 回数を設定
> myData <- replicate(mc,myTrial())
> ## 簡単な集計
> table(myData) # 頻度

myData
change stay
6673 3327

> table(myData)/mc # 確率 (推定値)

myData
change stay
0.6673 0.3327

```

5.2.4 St Petersburg のパラドックス

次の例は、無限回の試行を行う理論と、有限回しか実行できない現実との関係を考える問題である。

問題 5.4. 偏りのないコインを表が出るまで投げ続け、賞金を貰うゲームを考える。表が出るまでにコインを投げた回数が n 回であるとき、貰える賞金は 2^n 円とする。このとき賞金の期待値は

$$\mathbb{E}[\text{賞金}] = 2 \times \frac{1}{2} + 2^2 \times \frac{1}{2^2} + 2^3 \times \frac{1}{2^3} + \dots = \infty$$

となるが、ゲームを行う回数は現実には有限回であるが、それでも期待値は発散すると考えて良いであろうか？

ゲームを繰り返し無限回行う場合には上記の期待値の計算は正しいが、実際に無限回行うことはできない（例えば人には寿命がある）。有限回（例えば 100 回）でやめざるを得ないときに、実際の「期待値」はいくつになるか調べてみよう。

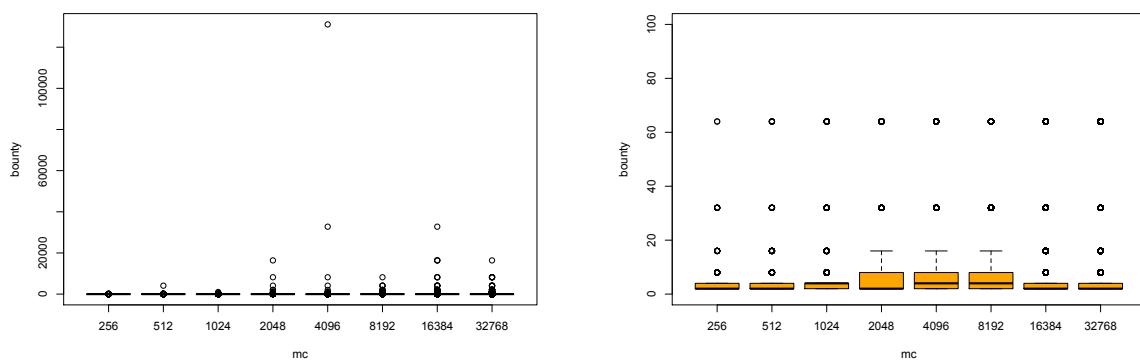
図 5.3 参照

```

> ### St Petersburg のパラドックス
> ## コイン投げの賭けの試行
> myTrial <- function(verbose=FALSE){
+   num <- 0 # コイン投げの回数
+   while(TRUE){ # 条件が満たされるまでループする

```

Rscript: [mc-stpetersburg.r](#)



(a) 実験結果の箱ひげ図

(b) 0 付近の様子を拡大した図

図 5.3: シミュレーションの例。

```

+      num <- num + 1 # 回数を増やす
+      if(runif(1)>0.5) break # 表が出たら終了
+      ## 一様分布で値が 0.5 を越えらる表として模擬
+      ## 以下の二項分布でも同じ
+      ## if(rbinom(1,size=1,prob=0.5)==1) break
+    }
+    bounty <- 2^num # 賞金を計算
+    if(verbose==TRUE){ # 何回投げたかも返す
+      return(c(num=num,bounty=bounty))
+    } else { # 賞金だけ返す
+      return(bounty)
+    }
+  }
> ## 試行を行ってみる
> myTrial()
[1] 2
> myTrial(verbose=TRUE)

  num   bounty
 1       2

> ## Monte-Carlo simulation
> ## set.seed(8888) # 実験を再現したい場合はシードを指定
> myDF <- data.frame(mc=NULL,bounty=NULL)
> for (mc in c(2^(8:15))) {
+   myData <- replicate(mc,myTrial())
+   cat("試行回数: ", mc, "\n")
+   cat("賞金の平均値: ", mean(myData), "\n")
+   myDF <- rbind(myDF,data.frame(mc=rep(mc,mc),
+                                 bounty=myData))
+ }

試行回数: 256
賞金の平均値: 6.515625
試行回数: 512
賞金の平均値: 15.89453
試行回数: 1024
賞金の平均値: 9.96875
試行回数: 2048
賞金の平均値: 25.46289
試行回数: 4096
賞金の平均値: 53.42041

```

```
試行回数: 8192
賞金の平均値: 13.41406
試行回数: 16384
賞金の平均値: 18.66077
試行回数: 32768
賞金の平均値: 13.76599
```

```
> boxplot(bounty ~ mc, data=myDF) # 回数ごとの箱ひげ図
> boxplot(bounty ~ mc, data=myDF,
+           ylim=c(0,100), col="orange") # y 軸を制限
```

5.2.5 秘書問題

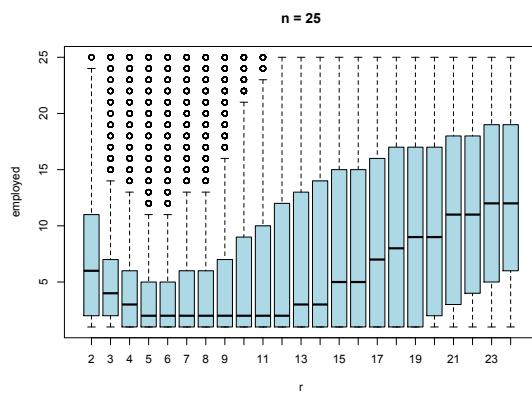
最後は、現実にありえる問題を単純化して数学的にも扱い易くしたものである。

問題 5.5. 秘書を 1 人雇いたいとする。前提条件は以下のとおりである。

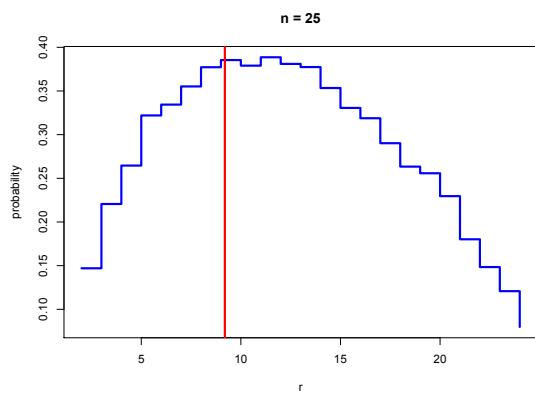
1. n 人が応募しており、 n は既知とする。
2. 応募者には 1 位から n 位まで同順位無しで順位付けできる。
3. 無作為な順序で 1 人ずつ面接を行う。
4. 毎回の面接後、その応募者を採用するか否かを決定する。
5. 不採用にした応募者を後から採用することはできない。

これに対して「面接者は最初の $r - 1$ 人の応募者をスキップし、その次の応募者がそれまで面接した中で最もよい応募者なら採用する」という戦略を取るとき、最良の応募者を採用する確率を最も高くするためには r をいくつとすれば良いか？

これは最適停止問題とよばれる最適戦略を問う問題の一種である。問題の条件によっていろいろな戦略が考えられているが、上記は最も単純なものである。



(a) r ごとの採用順位の分布



(b) r ごとの 1 位の採用確率

図 5.4: シミュレーションの例。

Rscript: mc-secretary.r

図 5.4 参照

```

> ### 秘書問題
> ## 秘書の採用の試行
> myTrial <- function(n,r,verbose=FALSE){ # n と r を指定
+   applicants <- sample(1:n,size=n)
+   ref <- applicants[1:(r-1)]
+   test <- applicants[r:n]
+   idx <- which(test < min(ref))
+   if(length(idx)==0) {
+     employed <- applicants[n]
+   } else {
+     employed <- test[idx[1]]
+   }
+   if(verbose==TRUE){ # 全順位も返す
+     return(list(applicants=applicants,
+                 employed=employed))
+   } else { # 採用した者の順位のみ返す
+     return(employed)
+   }
+ }
> ## 試行を行ってみる
> n <- 10
> myTrial(n,2,verbose=TRUE)

$applicants
[1] 2 3 10 5 8 7 6 1 4 9

$employed
[1] 1

> myTrial(n,3,verbose=TRUE)

$applicants
[1] 3 5 2 7 9 8 10 6 1 4

$employed
[1] 2

> myTrial(n,4,verbose=TRUE)

$applicants
[1] 1 9 7 2 3 4 10 5 6 8

$employed
[1] 8

> myTrial(n,5,verbose=TRUE)

$applicants
[1] 4 7 2 8 6 9 10 1 3 5

$employed
[1] 1

> myTrial(n,6,verbose=TRUE)

$applicants
[1] 4 6 5 1 2 8 3 7 10 9

$employed
[1] 9

```

```

> ## Monte-Carlo simulation
> ## set.seed(8888) # 実験を再現したい場合はシードを指定
> mc <- 5000
> n <- 25 # 候補者数を変えて実験
> myDF <- data.frame(r=NULL,employed=NULL)
> for (r in 2:(n-1)) {
+   myData <- replicate(mc,myTrial(n,r))
+   if(r %in% c(2,6,10,14,18,22)) { # いくつか表示
+     cat("試行回数: ", r, "\n")
+     print(table(myData))
+   }
+   myDF <- rbind(myDF,data.frame(r=rep(r,mc),
+                                   employed=myData))
+ }

試行回数:  2
myData
  1   2   3   4   5   6   7   8   9   10  11  12  13  14
735 549 444 381 352 299 292 259 240 191 179 155 150 137
  15  16  17  18  19  20  21  22  23  24  25
113 122 85  64  67  54  40  40  26  17  9

試行回数:  6
myData
  1   2   3   4   5   6   7   8   9   10  11
1672 911 559 370 251 198 141 81  63  71  62
  12  13  14  15  16  17  18  19  20  21  22
  49  44  42  35  52  40  37  49  49  44  54
  23  24  25
  46  39  41

試行回数:  10
myData
  1   2   3   4   5   6   7   8   9   10  11
1895 768 413 191 155 103 73  77  81  79  87
  12  13  14  15  16  17  18  19  20  21  22
  64  86  90  74  64  73  80  79  80  78  67
  23  24  25
  72  83  88

試行回数:  14
myData
  1   2   3   4   5   6   7   8   9   10  11
1767 547 230 144 127 125 127 97  122 113 112
  12  13  14  15  16  17  18  19  20  21  22
  105 104 108 105 81  115 113 106 122 98  124
  23  24  25
  105 91  112

試行回数:  18
myData
  1   2   3   4   5   6   7   8   9   10  11
1317 330 189 168 145 142 132 153 158 152 120
  12  13  14  15  16  17  18  19  20  21  22
  140 143 133 159 128 133 171 142 126 153 149
  23  24  25
  130 137 150

試行回数:  22
myData
  1   2   3   4   5   6   7   8   9   10  11  12  13  14
742 248 172 183 184 157 175 178 186 170 159 147 188 156
  15  16  17  18  19  20  21  22  23  24  25
  186 179 189 202 168 170 152 170 158 177 204

> boxplot(employed ~ r, data=myDF, # rごとの箱ひげ図
+           col="lightblue", main=paste("n =", n))

```

```

> (myDF2 <- aggregate(myDF[, "employed"],
+                      by=list(r=myDF$r),
+                      FUN=function(x){mean(x==1)}))

      r      x
1   2 0.1470
2   3 0.2206
3   4 0.2646
4   5 0.3220
5   6 0.3344
6   7 0.3552
7   8 0.3772
8   9 0.3854
9  10 0.3790
10 11 0.3886
11 12 0.3810
12 13 0.3774
13 14 0.3534
14 15 0.3306
15 16 0.3188
16 17 0.2902
17 18 0.2634
18 19 0.2558
19 20 0.2296
20 21 0.1802
21 22 0.1484
22 23 0.1208
23 24 0.0798

> plot(x ~ r, data=myDF2, # 1位を採用できる確率を表示
+       type="s", col="blue", lwd=3,
+       main=paste("n =", n), ylab="probability")
> ## 理論的に良いとされる r の値 (n が十分大きい場合)
> n/exp(1)

[1] 9.196986

> abline(v=n/exp(1), col="red", lwd=3)

```

5.3 補遺

5.3.1 参考文献

この章に関連する参考書としては以下を挙げておく。

- [1] 金明哲. *Rによるデータサイエンス(第2版)*. 東京: 森北出版, 2017.
- [2] U. リゲス (石田基広訳). *Rの基礎とプログラミング技法*. 東京: 丸善出版, 2012.
- [3] 奥村晴彦. *Rで楽しむ統計*. 東京: 共立出版, 2016.

索引

MASS, 26
data.table, 51
dplyr, ii, 34, 38, 45, 69
ggplot2, ii, 63, 64
graphics, 53
nycflights13, 51
readr, 51
scatterplot3d, 61
aes, 64
aggregate, 43
apply, 43
array, 31, 34
barplot, 58, 59
boxplot, 57, 58
c, 8, 34
cbind, 30
colnames, 30
colors, 53
curve, 53
data, 34, 51, 63
data.frame, 12, 32, 34
det, 19
diag, 19
dim, 30, 31
dplyr::arrange, 46
dplyr::desc, 46
dplyr::distinct, 48
dplyr::filter, 46
dplyr::group_by, 50
dplyr::mutate, 49
dplyr::rename, 47
dplyr::sample_frac, 50
dplyr::sample_n, 50
dplyr::select, 47
dplyr::slice, 46
dplyr::summarize, 49
dplyr::transmute, 49
eigen, 27
example, 53
fread, 51
function, 23
geom_bar, 70
geom_boxplot, 68
geom_histogram, 67
geom_line, 69
geom_point, 64
geom_smooth, 66
getwd, 39
ggplot2::geom_bar, 69
ggplot2::geom_boxplot, 67
ggplot2::geom_freqpoly, 66
ggplot2::geom_histogram, 66
ggplot2::geom_line, 68
ggplot2::geom_point, 64
ggplot2::geom_smooth, 65
ggplot2::ggplot, 64
ginv, 26, 28
help, 3, 34, 53
help.search, 3
hist, 56, 57
install.packages, 4
labs, 64
legend, 54, 55
length, 8, 30
lines, 53
list, 10
load, 41
matrix, 9, 30, 34
max, 24, 42
mean, 42
merge, 38
min, 42
mode, 29
names, 33
ncol, 30
norm, 25
nrow, 30
pairs, 60
par, 55, 62
persp, 61
pie, 59
plot, 53, 54, 60
points, 53
q, 2
rbind, 30
rbinom, 71, 73
read.csv, 39, 41
read.table, 51
rep, 8, 34
req, 30

`rnorm`, 14
`rownames`, 30, 33
`runif`, 14, 71, 73
`sample`, 14, 50, 71
`save`, 41, 42
`scatterplot3d`, 61
`seq`, 8, 30, 34
`set.seed`, 71
`setwd`, 39
`sin`, 53
`solve`, 20, 21
`split`, 38
`subset`, 33, 37, 45
`sum`, 19, 24, 42
`svd`, 27
`t`, 18
`tapply`, 44
`typeof`, 29
`write.csv`, 39, 41, 42
`write_csv`, 51

Comprehensive R Archive Network, 1
CRAN, 1

Monte-Carlo method, 71

object, 6

package, 1
pseudo random number, 71

R Project, 1
R Project for Statistical Computing, 1
random number, 71
RStudio, 1

stochastic simulation, 71

オブジェクト, 6
確率シミュレーション, 71
擬似乱数, 71
パッケージ, 1
モンテカルロ法, 71
乱数, 71