第 I 部

データ解析環境 R

R 入 門

1 2

R の基礎知識

本章では，R の生い立ちとデータ解析環境としての位置づけを与え，次に，R に関するマニュアルや参考文献，利用上有益となるインターネット上の情報を与える．さらに，R のインストールに際して留意すべき点や，コンピュータ言語・ソフトウェアとしての R の特徴にも言及する．これらの事項を学ぶことによって，「R とは何か」，「R を利用するための情報はどのように得ることができるのか」，「R を実際に使うための準備においてどのようなことに留意すべきか」という，非常に素朴かつ重要な R 自体に関する知識を体系的に得ることができるであろう．

1.1 2 3 4 R とは

check box □□□

R は，オープンソースのデータ解析環境 (data analysis environment) の一つである．現在は “The R Core Team” [1] を中心として保守や改良などが行われており，多様な分野やユーザ層に普及している．

データ解析環境としての R の位置づけを明確にするためには，**探索的データ解析** (Exploratory Data Analysis: EDA) の提唱と S 言語の開発についてふれることが自然であろう．まず，データ解析に関係する統計科学の分野における一つのブレークスルーは，J. W. Tukey (1977) によって提唱された探索的データ解析である．探索的データ解析は，モデルを構築する前にデータを数値的に**要約** (summarization) したり，図式的に**可視化** (visualization) することによってデータ自身のもつ情報を探索的に引き出し，その結果をふまえた**統計モデリング** (statistical modeling) を行う方法を提供する (図 1.1 を参照)．

探索的データ解析をコンセプトとし，それを具体的に実現するためのソフトウェア環境が R. A. Becker と J. M. Chambers によって S という名称で**設計** (design) され，S システム (後に S 言語) として**実装** (implementation) された (Becker and Chambers (1984) を参照)．

また R は，1990 年代のはじめに R. Gentleman と R. Ihaka によって開発が開始され[2]，S 言語の文法を参考にしながら発展してきた[3] (Ihaka (1998)

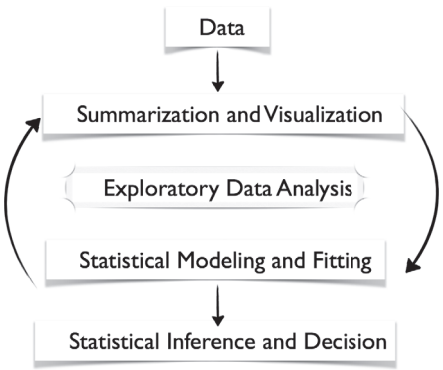

図 1.1 探索的データ解析 (EDA) に関する概念図

1) `https://www.r-project.org/`

2) R という名称は，開発者のファーストネームの頭文字が理由の一つといわれる．なお，R 開発の動機については，Ihaka (2009) に興味深い記述がある (`https://www.stat.auckland.ac.nz/~ihaka/downloads/Waikato.pdf`)．

3) R の開発は S に影響を受けたことは事実と思われるが，Ross Ihaka 氏によると，むしろ Scheme という Lisp 系の言語に強い影響を受けたとのことである．

を参照).

この一連の流れから，R は共通の設計にもとづく「S の異なった実装」とみることができよう (図 1.2 を参照)．なお，S, R の開発・発展に関する歴史的な背景については，柴田 (2015) に興味深い記述がある．また，最新の R 環境による探索的データ解析の実行については Wickham and Grolemund (2016) に与えられているので参照してほしい．

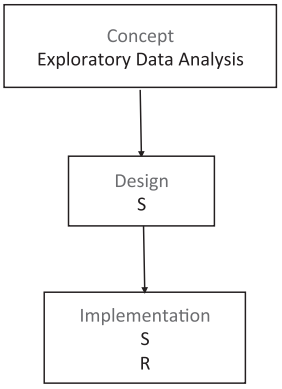

図 1.2 S, R 開発の流れ

1 1.2 3 4 Rの特徴 check box □□□

R の特徴を以下に与える．

(R1) 各種のオペレーティング・システム (Operating System: OS) に対応

Windows[4]，macOS[5]，Linux[6] などの主要な OS に対応したインストーラが利用できる[7]．

(R2) フリーソフトウェア

Free Software Foundation (FSF)[8] によって制定された GNU[9] General Public License[10] の意味でのフリーであり，R の再配布や改変などが可能である．

(R3) オープンソース

ソースコードが開示されており，R の設計や仕様を知ることが可能である．

(R4) オブジェクト指向プログラミング言語

R で扱われるすべての対象は**オブジェクト** (object) とよばれ，統一的に扱われる[11]．

(R5) 関数型言語

R での処理は**関数** (function) によって行われる．**演算子** (operator) も関数である．

---

4) Windows は，米国 Microsoft Corporation の米国およびその他の国における登録商標です．

5) macOS は，米国および他の国々で登録された Apple Inc. の商標です．

6) Linux は，Linus Torvalds 氏の日本およびその他の国における登録商標または商標です．

7) ソースコードも配布されているので，UNIX などでは `configure`, `make` コマンドを利用してユーザが構築することも可能である．

8) 1985 年 10 月 4 日に Richard Stallman によって創設された非営利団体である．当団体は，フリーソフトウェア運動，すなわち，コンピュータ・ソフトウェアを作成，頒布，改変する自由をユーザーに広く遍く推し進めることを狙い，コピーレフトを基本とする社会運動の支援を目標に掲げている (`https://www.fsf.org/`)．

9) Gnu is Not Unix の略．UNIX 互換のソフトウェア環境をすべてフリーソフトウェアで実装することを目的とし，FSF によって進められているプロジェクト (`https://www.gnu.org/`)．

10) 一般公衆利用許諾契約書のこと．GPL と略される．

11) 一般に，扱う対象をオブジェクトとして扱うプログラミングは，**オブジェクト指向プログラミング** (Object Oriented Programing: OOP) とよばれ，オブジェクト指向プログラミングが可能なコンピュータ言語を**オブジェクト指向プログラミング言語** (OOP languaege) という．R におけるオブジェクト指向に関する詳細は，Wickham (2014) を参照してほしい．

R の基礎知識

本章では，R を利用する上での必要最小限の知識を学ぶ．内容としては，R の起動・終了と，R によって作業を行う場所を指定する方法を与え，基本的なヘルプ機能について述べる．また，実数の四則演算や対数関数などの数学で利用される関数を R で利用する方法を与え，データ解析に利用する際に必ず学ばなければならないベクトルや行列，文字列，論理値，特殊値などに関する話題も扱う．これらの事項を学ぶことによって，R を利用する上での必要最小限の知識を得ることができる．

なお，本章で与えた R のスクリプト (コード) を実際に入力し，結果を確認しながら読み進めることによって，R の基本的な機能を知ることができると共に，「R との対話」がどのようなものかということを実体験できよう．ただし，R を利用する上で重要となるスクリプトの管理やグラフィック機能などの情報については付録 B を参照してほしい．

2.1 R の起動と終了

2 3 4 5 6 7 8 9

check box □□□

R を起動すると[1]，[R Console] (または [R コンソール]) にメッセージが出力された後，以下のような状態となる．

```
>
```

ここで，“>” は R **プロンプト** (R prompt) とよばれる．R プロンプトに続いてユーザが各種の命令を入力すると，その結果が R から返される[2]．例えば，R を終了する方法は，プロンプトに続いて `q()` と入力すればよい．

```
> q()
```

Windows 環境では，この入力によって，[作業スペースを保存させますか?] というメッセージがダイアログボックスに表示され，起動中に作成されたオブジェクトを保存するか否か，または終了をキャンセルするかどうかを問われるので，保存したい場合は [はい (Y)]，保存したくない場合は [いいえ (N)]，終了自体をキャンセルしたいときは，[キャンセル] を選択する．

また，macOS 環境では，[R セッション終了 ワークスペースのイメージファイルを保存しますか?] というメッセージが表示され，保存したい場合は，保存 を選択し，保存したくない場合は 保存しない を選択する．なお，終了自体をキャンセルしたい場合は キャンセル を選択する．

なお，保存を選択した場合は，結果が R の作業ディレクトリに `.RData` というファイル名で

1) Windows 環境下では，インストール後にデスクトップに作成された R のショートカットアイコンをダブルクリックすればよい．なお，Windows 10 では，[スタートメニュー] の [すべてのアプリ] にある R のフォルダにもアイコンが登録されている．また，macOS 環境では Launchpad にアイコンが登録されている．さらに，Ubuntu 環境では端末 (ターミナル) を起動後，`R` とタイプすればよい．

2) このような仕様の言語は「対話型言語」といわれる．

保存され，入力履歴が.Rhistory ファイル[3]に保存される．また，専用メニューから R を終了させることも可能である．

2.2 作業ディレクトリと作業空間

check box □□□

R に限らず，一般にソフトウェアを利用するときに，

「どの場所の何というファイルを利用しているのか?」

ということを意識することは重要である．最近のコンピュータ環境は検索機能が優れているため，ファイルの場所や名称を意識することが軽視されがちであることは否めないが，このことを意識していないと，以前に作ったファイルが見当たらず，探すのに思わぬ手間と時間をとられることがある．

R で作業を行う場所は**作業ディレクトリ** (working directory) とよばれ，作業を行った結果を保存するファイルを**作業空間**または**作業スペース** (workspace) という．

作業空間はデフォルト[4]では，R のホームディレクトリ[5]にある.RData というファイルである．

現在の作業ディレクトリを調べる方法は，関数 getwd を以下のように利用する．

```
> getwd()
```

Windows 環境下でこの関数による出力が，例えば C:/Rhome であったときは，C ドライブにある作業ディレクトリ Rhome への階層を示しており，/(スラッシュ) は階層構造の**分割符** (separator) を表す[6]．また，明示的に作業ディレクトリを変更するためには，関数 setwd を利用する．例えば，作業ディレクトリを C:/Rhome2 に変更したい場合は以下のように入力する．

```
> setwd("C:/Rhome2")
```

このような方法で変更した作業ディレクトリに作業空間.RData が存在し，それを利用して作業を行いたい場合は，以下のように関数 load を利用する．

```
> load(".RData")
```

なお，作業空間の名前はデフォルトでは.RData であるが，これ以外にも任意のファイル名が利用できる．ただし，あまり一般的な名称を用いてしまうと，そのファイルが R の作業空間であることの判別ができない可能性があるので，例えば，mywork.RData のように.RData をファイルの**拡張子** (extension) として利用すればよい．

3)　.Rhistory ファイルにはすべての入力履歴がテキストファイルとして保存されているため，この記録を適当なエディタで編集することによって，関数の作成や処理の自動化などに再利用することができる．

4)　コンピュータに関連する用語で**デフォルト** (default) とは，ユーザーが特に指定しない場合に設定されている標準の動作条件を指している．なお，引数に対しては，省略した場合の値を意味することに注意しよう．

5)　Windows の場合，スタートメニューなどに登録されたアイコンで R を起動したときのホームディレクトリは，各ユーザのドキュメントフォルダである．ただし，「日本語」を含むフォルダ (ディレクトリ) 名を利用している場合には R が正常に動作しない場合があるため，注意が必要である．半角英数文字 (ASCII(American National Standard Code for Information Interchange) コード) のみのフォルダ (ディレクトリ) 名が強く推奨される．

6)　通常，Windows における分割符はバックスラッシュ (\) であるが，R ではスラッシュ (/) で表し，日本語版 Windows 環境下における分割符は円マーク (¥) であることに注意してほしい．

第 II 部

R による統計学の基礎

確率変数と確率分布

5 Rにおける関数の定義
6
7 多変量確率変数と多変量確率分布

改めて我々の身の回りを見渡すと，身長や体重等の身体的な値はもとより，年間の平均気温や降水量等の気象条件，企業の年間の売上高や資産合計などの財務指標は個人・年・企業ごとにランダムに変動をしており，むしろ確定的なものを探すことの方が難しい．統計学では，このようなランダムな現象を確率をともなう変数，すなわち「確率変数」として扱い，確率変数のとる値の分布法則を「確率分布」で規定する．このことは，確率分布を「データ発生メカニズム」と見なす，と言い換えることができる．

本章では，確率変数と確率分布の定義を与え，代表的な確率分布の諸性質を R を使って確かめる．実際に R のコードを実行し，対話的に学ぶことによって，確率分布を理論的・数値的に理解することができ，データを擬似的に生成することが可能となる．なお，解説の都合上，新しく必要となった関数を付録 E.2 に与えた[1)]．

6.1 統計学における確率の役割

2 3 4 5 6 check box □□□

統計学は，**母集団** (population) に対して**調査** (survey)，**実験** (experiment)，**観測** (observation) といった**試行** (trial) を行い，その**結果** (outcome) の集まり (**事象** (event) とよばれる) から得られる情報に基づいて母集団に対する**推測** (inference) や**決定** (decision) を行うための理論や手法を与える．その際，事象の起こりやすさを**確率** (probability) を用いて $[0, 1]$ の範囲内の値で数値化することによって評価する (確率の定義の詳細は付録 A.2 を参照)．ここで，母集団と結果を Ω と ω を用いて表し，事象をアルファベットの大文字 A, B, C などを用いて表す．また，事象 A の確率を $P(A)$ と書く．

6.2 確率変数

1 3 4 5 6 check box □□□

具体的な例として，日本における新生児の体重を調査することを考えよう．まず，調査の対象となる新生児の全体が母集団 Ω であり，一人の新生児が結果 ω となる．また，体重に関するある種の条件を満たす新生児の集まりが事象 A である．調査は新生児の体重を観測することによって行われるので，体重を X などの記号で明記すると便利である．ここで，体重 X は新生児 ω に依存するため，ω の関数 $X = X(\omega)$ と考えることができる．また，新生児全体に対して体重がある範囲内に入る新生児の割合を考えると，体重は確率 (割合) をともなう変数と考えることができる．このように変数が確率をともなうと考えることができる場合，一般に，**確率変数** (random variable) とよばれる．

1) 付録 E.2 に与えた関数のソースコードのファイルは「本書の使い方」で述べた方法で入手可能である．

例えば，「体重 X が 2500g から 3500g の新生児の集まり」という事象を

$$\begin{aligned} A &= \{\text{ 体重が 2500g から 3500g の新生児の集まり }\} \\ &= \{\omega \in \Omega \mid 2500 < X(\omega) < 3500\} \end{aligned}$$

とおいたとき，その割合 (確率) は，

$$P(A) = P\left(\{\omega \in \Omega \mid 2500 < X(\omega) < 3500\}\right) = P(2500 < X < 3500)$$

で表される．一般に，確率変数はアルファベットの大文字 X, Y, Z などを用いて表されることがある．なお，統計学では**変量** (variate) ともよばれることに注意しよう．

1 2 **6.3** 4 5 6

確率分布

check box □□□

6.3.1 分布関数

確率変数 X は確率をともなって値をとるが，それを規定するものが**確率分布** (probability distribution) P^X である[2)]．確率分布は**分布法則**，または単に，**分布** (distribution) ともよばれることがある．確率変数 X の確率分布が P^X であるとき，「確率変数 X は確率分布 P^X に従う」とよばれ，記号として $X \sim P^X$ と書く．

確率分布を具体的に表現することを考える．まず，確率変数 X が x 以下の値をとる確率を

$$F(x) := P(X \leq x) \tag{6.1}$$

と表し，**分布関数** (distribution function) または**累積分布関数** (cumulative distribution function) とよぶ．ある点 x_0 が固定されたときの分布関数の値 $F(x_0)$ は確率変数 X が $(-\infty, x_0]$ の範囲をとる確率を表しており，この確率は点 x_0 の**下側確率**[3)] とよばれる．

分布関数を利用すると，確率変数 X が任意の区間の値をとる確率を計算することができる．この意味で，確率分布を規定することに注意しよう．例えば，任意の区間を $(a, b]$ とすると，

$$P(a < X \leq b) = P(X \leq b) - P(X \leq a) = F(b) - F(a) \tag{6.2}$$

となる．

分布関数の性質としては，以下の 3 つのものがある．

(DF1) 単調非減少性： $x < y \implies F(x) \leq F(y)$

(DF2) 右連続性[4)]： $\lim_{x \to a+0} F(x) = F(a)$

(DF3) $0 \leq F(x) \leq 1, \quad \lim_{x \to -\infty} F(x) = 0, \quad \lim_{x \to \infty} F(x) = 1$

---

2) より正確には，確率変数 X によって確率空間 $(\Omega, \mathcal{A}, P)$ から可測空間 $(\mathbb{R}, \mathcal{B})$ へ誘導される確率測度 P^X が確率分布とよばれる．

3) 「下側」を「左側」，「上側」を「右側」とよぶこともある．

4) ここで，$\lim_{x \to a+0} f(x)$ は x が右側から点 a へ近づいたときの関数 $f(x)$ の極限 (右極限) を表す記号である．

推定・検定を行う際に，正規分布 $N(\mu, \sigma^2)$ に従う確率変数 X が区間 $(a, b]$ に値をとる確率 (面積) を求めたり，逆に，指定された確率と同等の**裾** (tail) の面積を与える点 (分位点，パーセント点) を計算する必要が生じる．ここでは，これらの値を R を用いて求める方法を扱う．なお，ここで扱われる事項と関連して付録 C.1 も参照してほしい．

まず，確率変数 X が区間[15] $(a, b]$ の範囲に値をとる確率は，一般に，分布関数を利用すると，

$$P\left(a < X \leq b\right) = F(b) - F(a)$$

で求めることができることを思い出そう ((6.2) 式も参照)．この結果を使って，正規分布 $N(0, 2^2)$ に従う確率変数 X が区間 $(-1, 3]$ に値をとる確率を R を用いて計算すると，

```
> pnorm(3,0,2)-pnorm(-1,0,2)
[1] 0.6246553
```

となる．この結果は，正規分布 $N(0, 2^2)$ の分布関数を $F(x)$ としたとき，

$$\begin{aligned} P(-1 < X \leq 3) &= F(3) - F(-1) \\ &= 0.9331928 - 0.3085375 \\ &= 0.6246553 \end{aligned}$$

であることを表している．

なお，図 6.5 には，確率変数 X が区間 $(-1, 3]$ に値をとる確率を表す領域が描かれており，実際にこの図を描くためには付録 E.2 に与えた関数 `ggplot.pdf.normal.area` を以下のように利用する．

```
> ggplot.pdf.normal.area(lb=-1,ub=3,mean=0,sd=2)
```

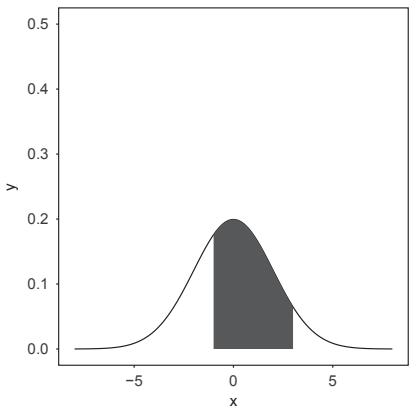

図 6.5　密度関数 (正規分布 $N(0, 2^2)$) の区間 $(-1, 3]$ の領域

次に，与えられた正規分布の裾の面積に対する分位点を，関数 `qnorm` を用いて求めることを考える．簡単のため，標準正規分布の場合を考え，下側確率が 5% となる点 (下側 5% 点または 5% 分位点) を求めるためには以下のように入力すればよい．

15)　正規分布は連続型密度をもつ分布であるため，確率変数 X が任意の点 x をとる確率は 0 となり，区間が $(a, b]$, $[a, b)$, $[a, b]$, (a, b) のどのパターンでも，その範囲の確率はすべて一致することに注意しよう．

```
> qnorm(0.05)
[1] -1.644854
```

なお，図 6.6 には，下側 5% 点以下の領域が描かれており，実際にこの図を描くためには，付録 E.2 に与えた関数 ggplot.pdf.normal.area を以下のように利用する．

```
> ggplot.pdf.normal.area(lb=-4,ub=qnorm(0.05))
```

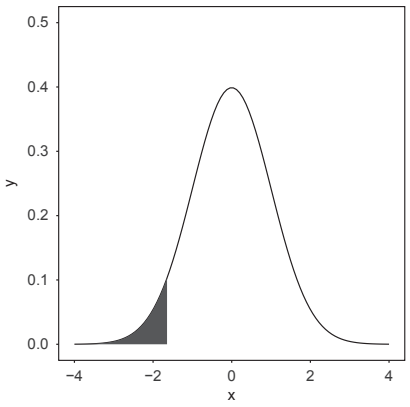

図 6.6 密度関数 (標準正規分布) の下側 5% の領域

逆に，上側確率が 5% となる点 (上側 5% 点) は下側 95% 点 (95% 分位点) と一致し，以下のように入力することによって求めることができる．

```
> qnorm(0.95)
[1] 1.644854
```

より直接的な方法としては，関数 qnorm の引数 lower.tail に FALSE を与えることによって上側 5% 点を計算することができる．

```
> qnorm(0.05,lower.tail=FALSE)
[1] 1.644854
```

ここで，lower.tail は下側の裾 (lower tail) を計算するかどうかを判断する引数であり，FALSE を与えることによって，下側の裾ではなく上側の裾の確率を計算することが可能となる．なお，図 6.7 には，上側 5% 点以上の領域が描かれており，下側 5% 点の場合と同様に，関数 ggplot.pdf.normal.area を使って以下のように利用する．

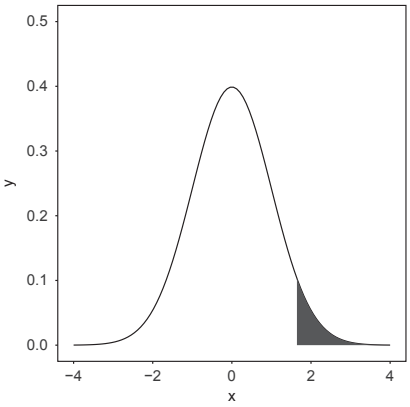

図 6.7 密度関数 (標準正規分布) の上側 5% の領域

1 2 **7.3** 4 5 6 7 8 9

2 変量連続型確率分布 check box □□□

2 変量正規分布

正規分布 $N(\mu, \sigma^2)$ は 1 変量連続型確率分布の最も代表的なものの一つであるが，その 2 変量への拡張である **2 変量正規分布** (bivariate normal distribution) も，また 2 変量連続型確率分布の代表的なものである．

2 変量正規分布は $N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ という記号で表される．ここで，

$$\boldsymbol{\mu} := \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \qquad \boldsymbol{\Sigma} := \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix} \tag{7.24}$$

である．なお，$\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$ は，この時点においてどのような役割を示す母数 (ベクトル，行列) であるかはわからないことに注意しよう．

親と子の身長や，新生児の身長と体重などは，2 変量正規分布に従う例としてしばしば取り上げられる．2 変量確率変数 (X, Y) が 2 変量正規分布 $N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ に従うときは $(X, Y) \sim N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ と表し，2 変量確率変数が $\boldsymbol{X} := [X, Y]'$ とベクトルで表現されているときは，$\boldsymbol{X} \sim N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ とも表す．

2 変量正規分布 $N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ の同時密度関数，周辺密度関数，同時分布関数，平均ベクトル，分散共分散行列は，それぞれ以下で与えられる．

同時密度関数:

$$f_{XY}(x, y) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2}Q(x, y)\right\} \tag{7.25}$$

ここで，

$$Q(x, y) := \frac{1}{1-\rho^2}\left\{\left(\frac{x-\mu_1}{\sigma_1}\right)^2 - 2\rho\left(\frac{x-\mu_1}{\sigma_1}\right)\left(\frac{y-\mu_2}{\sigma_2}\right) + \left(\frac{y-\mu_2}{\sigma_2}\right)^2\right\} \tag{7.26}$$

であり，

$$(x, y) \in \mathbb{R}^2, \qquad (\mu_1, \mu_2) \in \mathbb{R}^2, \qquad \sigma_1, \sigma_2 \in \mathbb{R}^+, \qquad \rho \in (-1, 1)$$

である．

周辺密度関数:

$$f_X(x) = \int_{-\infty}^{\infty} f_{XY}(x, y)\,\mathrm{d}y = \frac{1}{\sqrt{2\pi\sigma_1}} \exp\left\{-\frac{(x-\mu_1)^2}{2\sigma_1^2}\right\} \tag{7.27}$$

$$f_Y(y) = \int_{-\infty}^{\infty} f_{XY}(x, y)\,\mathrm{d}x = \frac{1}{\sqrt{2\pi\sigma_2}} \exp\left\{-\frac{(y-\mu_2)^2}{2\sigma_2^2}\right\} \tag{7.28}$$

同時分布関数:

$$F_{XY}(x, y) = \int_{-\infty}^{y}\int_{-\infty}^{x} f_{XY}(u, v)\,\mathrm{d}u\,\mathrm{d}v \tag{7.29}$$

平均ベクトル:

$$\mathrm{E}(\boldsymbol{X}) := \begin{bmatrix} \mathrm{E}(X) \\ \mathrm{E}(Y) \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} = \boldsymbol{\mu} \tag{7.30}$$

分散共分散行列:

$$\mathrm{V}(\boldsymbol{X}) := \begin{bmatrix} \mathrm{V}(X) & \mathrm{Cov}(X,Y) \\ \mathrm{Cov}(Y,X) & \mathrm{V}(Y) \end{bmatrix} = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_2\sigma_1 & \sigma_2^2 \end{bmatrix} = \boldsymbol{\Sigma} \tag{7.31}$$

以上の結果から，2 変量正規分布 $N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ に従う確率変数 (X, Y) のそれぞれの周辺分布は (1 変量) 正規分布 $N(\mu_1, \sigma_1^2)$, $N(\mu_2, \sigma_2^2)$ であり，$\boldsymbol{\mu}$ は平均ベクトル，$\boldsymbol{\Sigma}$ は分散共分散行列であることがわかる．また，

$$\mathrm{Cor}(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{V}(X)\,\mathrm{V}(Y)}} = \frac{\rho\sigma_1\sigma_2}{\sqrt{\sigma_1^2\sigma_2^2}} = \rho \tag{7.32}$$

となり，ρ は X と Y の相関係数である．なお，$\mathrm{Cov}(X,Y) = \mathrm{Cov}(Y,X) = \rho\sigma_1\sigma_2$ である．また，これらの結果は同時密度関数から直接導かれる結果である．

2 変量正規分布 $N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ において，相関係数 ρ が 0 のとき，すなわち無相関のとき，同時密度関数が,

$$\begin{aligned} f_{XY}(x,y) &= \frac{1}{\sqrt{2\pi}\sigma_1} \exp\left\{ -\frac{1}{2}\left(\frac{x-\mu_1}{\sigma_1} \right)^2 \right\} \frac{1}{\sqrt{2\pi}\sigma_2} \exp\left\{ -\frac{1}{2}\left(\frac{y-\mu_2}{\sigma_2} \right)^2 \right\} \\ &= f_X(x)\, f_Y(y) \end{aligned} \tag{7.33}$$

となる．このことは，2 変量確率変数 (X, Y) が相関係数 $\rho = 0$ の 2 変量正規分布に従うとき，確率変数 X, Y は，独立に正規分布 $N(\mu_1, \sigma_1^2)$, $N(\mu_2, \sigma_2^2)$ にそれぞれ従うことを示している．よって，相関係数の性質 (Cor2) の逆がいえて，2 変量正規分布に関しては以下の同値性が成り立つ．

$$X \perp\!\!\!\perp Y \iff \mathrm{Cor}(X,Y) = 0 \tag{7.34}$$

確率変数 X と Y が独立に正規分布 $N(\mu_1, \sigma_1^2)$, $N(\mu_2, \sigma_2^2)$ に従うとき，それらの和 $X + Y$ の分布は再び正規分布 $N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$ に従う．

$$X \sim N(\mu_1, \sigma_1^2), \quad Y \sim N(\mu_1, \sigma_2^2), \quad X \perp\!\!\!\perp Y \implies X + Y \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2) \tag{7.35}$$

この性質は，正規分布の**再生性** (reproductive property) とよばれる．

1 変量正規分布 $N(\mu, \sigma^2)$ において $(\mu, \sigma^2) = (0, 1)$ の場合は標準正規分布 $N(0, 1)$ とよばれたが，2 変量正規分布 $N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ において,

$$\mu_1 = \mu_2 = 0, \qquad \sigma_1 = \sigma_2 = 1, \qquad \rho = 0$$

のとき，すなわち,

$$\boldsymbol{\mu} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} = \mathbf{0}, \qquad \boldsymbol{\Sigma} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = \mathbf{I}_2 \tag{7.36}$$

のとき，2 変量標準正規分布 $N_2(\mathbf{0}, \mathbf{I}_2)$ とよばれる．

2 変量標準正規分布の同時密度関数は,

$$\begin{aligned}\phi_{XY}(x,y) &:= \frac{1}{2\pi}\exp\left(-\frac{x^2+y^2}{2}\right)\\ &= \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{x^2}{2}\right)\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{y^2}{2}\right)\\ &= \phi(x)\,\phi(y) \qquad (7.37)\end{aligned}$$

と表すことができる．ここで，$\phi(\cdot)$ は標準正規分布 $N(0,1)$ の密度関数である．このことは，2 変量確率変数 (X,Y) が 2 変量標準正規分布に従う場合に，確率変数 X, Y はそれぞれ独立に同一の標準正規分布 $N(0,1)$ に従うことを表している．

2 変量標準正規分布の同時密度関数と同時分布関数は，それぞれ図 7.1, 7.2 のように与えられる．

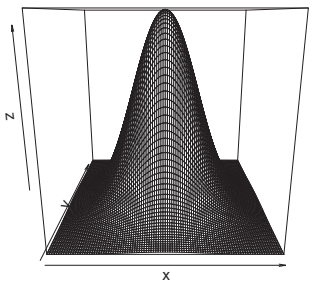

図 7.1 2 変量標準正規分布 $N_2(\mathbf{0},\mathbf{I}_2)$ の同時密度関数

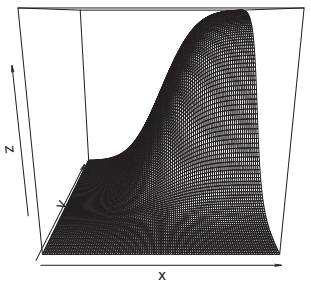

図 7.2 2 変量標準正規分布 $N_2(\mathbf{0},\mathbf{I}_2)$ の同時分布関数

これらの図からも 2 変量標準正規分布の同時密度関数と同時分布関数の概形を知ることができるが，特に同時密度関数は「視点」を動的に変化させることができなければ，一部の見えない部分が存在してしまう．この問題に対する一つの解決策は，**等高線** (contour line) をプロットすることによって，「山の高さ」を 2 次元でとらえることである．図 7.3 は，2 変量標準正規分布の同時密度関数の等高線プロットである．この図から，等高線は原点 $(0,0)$ を中心とする同心円で与えられることがわかる．

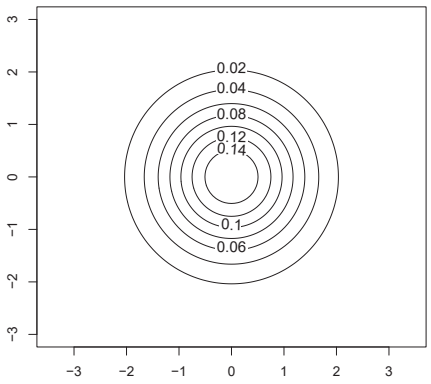

図 7.3 2 変量標準正規分布 $N_2(\mathbf{0},\mathbf{I}_2)$ の同時密度関数の等高線

2 変量正規分布を特別な場合に含む多変量正規分布を扱う関数は標準的なパッケージにはなく，追加パッケージ `mvtnorm` を導入することによって，以下のような R 関数を利用することができる[4)]．

表 7.1 `mvtnorm` パッケージに含まれる多変量正規分布に関する R 関数

関　数	役　割
`dmvnorm`	同時密度関数の値を求める．
`pmvnorm`	同時分布関数の値を求める．
`qmvnorm`	分位点を求める．
`rmvnorm`	乱数を発生させる．

4) 追加パッケージのインストールについては付録 B.6 を参照してほしい．

母集団分布と標本分布

7 多変量確率変数と多変量確率分布
8
9 データの要約と可視化

第 6，7 章では，ランダムな現象のメカニズムを確率変数と確率分布を導入することによって解説した．本章では，この考え方を援用して，調査や観測，実験の対象となる母集団から抽出された標本を確率変数と見なし，その確率分布を「母集団分布」とよぶ．さらに，無作為に抽出された標本にもとづく平均や分散を「統計量」とよび，その分布を「標本分布」と定義して明確に区別する．このことにより，標本平均などの統計量の変動が標本分布を用いて理論的・数値的に評価できるようになり，後の章で学ぶ母集団の平均 (母平均) や分散 (母分散) などに対する推定・検定を行う際の根拠となる．

本章では，母集団分布が正規分布に従う場合の標本平均の分布に関する性質を R を用いて可視化することによって理解し，さらに一般の母集団分布の場合についても漸近的な考察として，統計学における 2 大定理である「大数の法則」と「中心極限定理」を，シミュレーションを実行することによって直観的に理解する方法を試みる．また，正規分布から導かれる標本分布の密度関数や分位点関数の計算などを R を用いて計算する方法も学ぶ．

これらのことを R を利用しながら学ぶことが，統計学の初学者にとって難解と思われる標本分布という障壁を乗り越える一助となることを期待する．なお，解説の都合上必要となり新たに作成された関数を付録 E.4 に与えた[1)]．

8.1 母集団分布と標本分布

2 3 4 check box □□□

本章では，母集団 Ω に対する試行 (例えば，調査・観測・実験など) の結果 ω に対する変量 (確率変数) $X = X(\omega)$ を**標本** (sample) とよび，その分布 P^X を**母集団分布** (population distribution) という．標本 X が母集団分布 P^X に従うとき，$X \sim P^X$ と表される．

また，独立に同一の母集団分布 P^X に従う標本 $\{X_1, \cdots, X_n\}$ を**無作為標本** (random sample) とよび，

$$\{X_1, \cdots, X_n\} \overset{\text{i.i.d.}}{\sim} P^X \tag{8.1}$$

と表す．ここで，n は**標本の大きさ** (sample size) とよばれる．また，“i.i.d.” は，“**i**ndependent and **i**dentically **d**istributed”(独立に同一の分布に従う) を表す記号であることを思い出そう (7.9 節も参照)．

標本 X の (母集団) 分布 P^X が正規分布 $N(\mu, \sigma^2)$ であるとき，母平均 μ，母分散 σ^2 の**正規母集団** (normal population) とよばれる．正規母集団から得られた大きさ n の無作為標本 $\{X_1, \cdots, X_n\}$ は

$$\{X_1, \cdots, X_n\} \overset{\text{i.i.d.}}{\sim} N(\mu, \sigma^2) \tag{8.2}$$

と表される．

1) 付録 E.4 に与えた関数のソースコードのファイルは，「本書の使い方」で述べた方法で入手可能である．

さらに，実際に試行を行った結果に対する無作為標本 $\{X_1, \cdots, X_n\}$ の実現値 $\{x_1, \cdots, x_n\}$ を**データ** (data) という．

具体的な例としては，ある年の日本人の新生児の全体を母集団 Ω と考えると，結果 ω は (1 人の) 新生児[2]であり，その新生児の体重 $X = X(\omega)$ が標本となる．さらに，無作為に選ばれる (予定の) n 人の新生児の体重 $\{X_1, \cdots, X_n\}$ が無作為標本であり，実際に選ばれた n 人分の新生児の体重の値がデータ $\{x_1, \cdots, x_n\}$ となる．第 3 章で読み込まれたオブジェクト `weight` はデータの具体例である．なお，新生児の体重の母集団分布としては正規分布がしばしば利用される．

母集団分布が，連続型のとき密度関数 $f(x)$，離散型のとき確率関数 $p(x)$ をもつと仮定する．このとき，標本 X の平均は**母平均** (population mean) とよばれ，

$$\mu := \mathrm{E}(X) = \begin{cases} \displaystyle\int_{-\infty}^{\infty} x\, f(x)\, \mathrm{d}x & \text{(連続型)} \\ \displaystyle\sum_{x=-\infty}^{\infty} x\, p(x) & \text{(離散型)} \end{cases} \tag{8.3}$$

で定義される．母平均は母集団分布の位置の特性を表す値として利用される．また，標本 X の分散は**母分散** (population variance) とよばれ，

$$\sigma^2 := \mathrm{V}(X) = \mathrm{E}\left(X - \mathrm{E}(X)\right)^2 = \begin{cases} \displaystyle\int_{-\infty}^{\infty} \left\{x - \mathrm{E}(X)\right\}^2 f(x)\, \mathrm{d}x & \text{(連続型)} \\ \displaystyle\sum_{x=-\infty}^{\infty} \left\{x - \mathrm{E}(X)\right\}^2 p(x) & \text{(離散型)} \end{cases} \tag{8.4}$$

で定義される．母分散は母集団の拡がりの特性を表す．

母平均 μ や母分散 σ^2 などの母集団分布の特性値は，一般には，**母数**または**パラメータ** (parameter) とよばれ，θ という記号で表される．通常，母数 θ は未知であり，この値を統計的に推測 (推定，検定) するときに無作為標本 $\{X_1, \cdots, X_n\}$ の関数 $T_n := T(X_1, \cdots, X_n)$ が利用される．この関数 T_n は**統計量** (statistics) とよばれ，代表的なものとして，

標本平均 (sample mean)：

$$\overline{X}_n := \frac{1}{n}\left(X_1 + \cdots + X_n\right) = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{8.5}$$

標本分散 (sample variance)：

$$S_n^2 := \frac{1}{n}\left\{(X_1 - \overline{X}_n)^2 + \cdots + (X_n - \overline{X}_n)^2\right\} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2 \tag{8.6}$$

不偏分散 (unbiased variance)：

$$U_n^2 := \frac{1}{n-1}\left\{(X_1 - \overline{X}_n)^2 + \cdots + (X_n - \overline{X}_n)^2\right\} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2 \tag{8.7}$$

などがある．

母集団と標本に関する対応関係から，標本平均 $\overline{X}_n$，標本分散 S_n^2 は，それぞれ母平均 μ と母分散 σ^2 の推測に利用されることが直観的に理解できるが，そのことの理論的な検証は統計

2)　調査等を実施する前の状態で，具体的には決まっていないことに注意しよう．

的推定と統計的仮説検定を扱う第 10 章と第 11 章において行う．統計量 T_n は無作為標本 (確率変数) の関数であるので，やはり確率変数であり，ある種の分布法則に従うが，それを**標本分布** (sampling distribution) という．統計量の標本分布は, 推定や検定の精度を評価するために重要となる．なお, 表 8.1 に, 母集団分布と標本分布に関する記号と用語をまとめたので参照してほしい．

表 8.1 記号と用語

記　号	用　語	説　明
Ω	母集団	試行 (調査・観測・実験) の対象全体
ω	結果	試行の結果
$X := X(\omega)$	標本，観測	試行の結果 ω に対する変量 (確率変数)
P^X	母集団分布	標本 X の確率分布
$X \sim P^X$	標本 X は分布 P^X に従う	標本 X の確率分布が P^X であること
$\{X_1, \cdots, X_n\} \overset{\text{i.i.d.}}{\sim} P^X$	無作為標本	標本 $\{X_1, \cdots, X_n\}$ が独立に同一の分布 P^X に従うこと
$\{x_1, \cdots, x_n\}$	データ	独立な試行の結果に対して具体的に得られた無作為標本の実現値
θ	母数	母集団分布の特性値
μ	母平均	母集団分布の平均
σ^2	母分散	母集団分布の分散
$T_n = T(X_1, \cdots, X_n)$	統計量	無作為標本の関数
$\overline{X}_n$	標本平均	無作為標本の平均
S_n^2	標本分散	無作為標本の分散
U_n^2	不偏分散	無作為標本の不偏分散

次節では，具体的な統計量の標本分布として，標本平均の分布を R を用いて検証する．

1 **8.2** 3 4

標本平均の分布

check box □□□

8.2.1 正規母集団の場合

$\{X_1, \cdots, X_n\}$ を母平均 μ，母分散 σ^2 の正規母集団からの無作為標本，すなわち，$\{X_1, \cdots, X_n\} \overset{\text{i.i.d.}}{\sim} N(\mu, \sigma^2)$ とすると，標本平均 $\overline{X}_n$ に関して以下の命題 (性質) が成り立つ．

(SMN-1)

$$\mathrm{E}(\overline{X}_n) = \mu, \qquad \mathrm{V}(\overline{X}_n) = \frac{\sigma^2}{n}$$

(SMN-2)

$$\overline{X}_n \xrightarrow{P} \mu \quad (n \to \infty) \quad \overset{\text{def.}}{\Longleftrightarrow} \quad \forall \varepsilon > 0, \quad \lim_{n \to \infty} P(|\overline{X}_n - \mu| \geq \varepsilon) = 0$$

(SMN-3)

$$\overline{X}_n \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

第 III 部

R によるデータ解析と統計的推測

データの要約と可視化

8 母集団分布と標本分布
9
10 推　定

一般に，データは数字や記号の集まりであって，眺めているだけでは得られる情報は少ない．本章では，R を利用してデータの平均や分散などの**特性値** (characteristic value) または**要約値** (summarized value) を求めることによって**要約** (summarization) したり，データをプロットし**統計グラフィックス** (statistical graphics) を作成することによって**可視化** (visualization) し，データがもつ分布や相関関係などに関する情報を効率的に得るための方法について学ぶ．データを要約し可視化することは，**探索的データ解析** (Explanatory Data Analysis: EDA) を実践していることであり，まさにこのことを体得することが本章の目的である[1)]．

9.1 1 変量データの要約

2 3 4 5 6　check box □□□

R にデータが既に読み込まれているならば，関数 `summary` を用いて要約することができる．ここでは，3.1 節で読み込んだ新生児の体重のデータを実際に要約してみる．

```
> summary(weight)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   2270    2898    3160    3160    3368    4180
```

この結果において得られるのは，6 個の要約値[2)]であり，以下のようなものである．

`Min.` データの**最小値** (minimun) $x_{\min}$

`1st Qu.` データの**第 1 四分位点** (first quartile) Q_1

`Median` データの**中央値** (median) x_{med}

`Mean` データの**平均値** (mean) $\overline{x}$

`3rd Qu.` データの**第 3 四分位点** (third quartile) Q_3

`Max.` データの**最大値** (maximum) $x_{\max}$

各定義を以下に与える．

まず，平均値は，

$$\overline{x} := \frac{1}{n}\sum_{i=1}^{n} x_i := \frac{1}{n}(x_1 + \cdots + x_n) \tag{9.1}$$

1) 最新の R 環境のもとでの探索的データ解析の実行については Wickham and Grolemund (2016) も参照してほしい．

2) これらの要約値のうち平均値を除いたものは，**テューキーの 5 点要約値** (Tukey's five number summary) とよばれることがある．

で定義される．この値は，データ $\{x_1, \cdots, x_n\}$ 上でそれぞれ等質の「重み」$\frac{1}{n}$ をもつと考えたときに，それらの**重心** (center of gravity) という解釈ができ，この意味で，平均値はデータの (中心) **位置** (location) を表す特性値と考えられる．

平均値は関数 `mean` でも以下のように求めることができる．

```
> mean(weight)
[1] 3160.2
```

次に，中央値はデータの個数 n が奇数，偶数の場合に応じて以下のように定義される．

$$x_{\mathrm{med}} := \begin{cases} x_{(m)} & (n\text{: 奇数}) \\ \dfrac{x_{(m)} + x_{(m+1)}}{2} & (n\text{: 偶数}) \end{cases} \tag{9.2}$$

ここで，

$$m := \begin{cases} \dfrac{n+1}{2} & (n\text{: 奇数}) \\ \dfrac{n}{2} & (n\text{: 偶数}) \end{cases} \tag{9.3}$$

であり，

$$x_{(1)} \leq \cdots \leq x_{(n)} \tag{9.4}$$

は，データ $\{x_1, \cdots, x_n\}$ を昇順に並べ替えてできる**順序統計量** (order statistics) である．

中央値は，データを「半分」に分ける点であるという意味で中心位置を表す特性値であり，関数 `median` を用いても単独に求めることができる．

```
> median(weight)
[1] 3160
```

新生児の体重データにおける平均値 $\overline{x} = 3160.2$ と中央値 $x_{\mathrm{med}} = 3160$ はほぼ等しく，このような場合は，データは平均値 (または中央値) を中心としてほぼ**左右対称** (symmetric) に分布する．なお，これらの値がずれているときは左右どちらかに分布が偏っており，**歪み** (skewness) が生じている．

最小値と最大値は，順序統計量を使って，

$$x_{\min} = x_{(1)}, \qquad x_{\max} = x_{(n)}$$

と表せる．R では，関数 `min`, `max` を使って以下のように最小値と最大値を求めることができる．

```
> min(weight)
[1] 2270
> max(weight)
[1] 4180
```

関数 `which.min`, `which.max` を使うと，何番目のデータが最小値と最大値を与えるかを調べることができる．

```
> which.min(weight)
[1] 23
> which.max(weight)
[1] 100
```

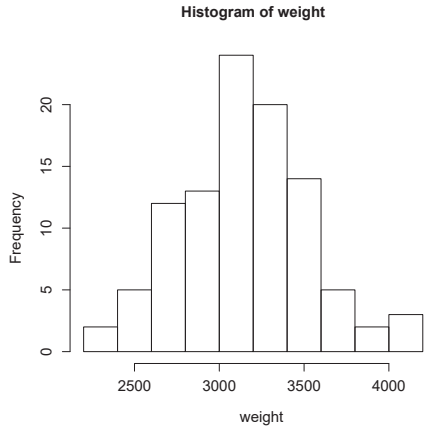

図 9.1　新生児の体重データのヒストグラム

級幅 200 のヒストグラムを描くためには，以下のように入力する．

```
> hist(weight,breaks=seq(2100,4300,200))
```

この入力によって，図 9.2 で与えられるヒストグラムが描かれる．図 9.1, 9.2 を比較すると，単峰であることには変化がないものの，形状が若干変化していることがわかる．一般に，ヒストグラムの「形状」は階級のとり方に強く依存する．

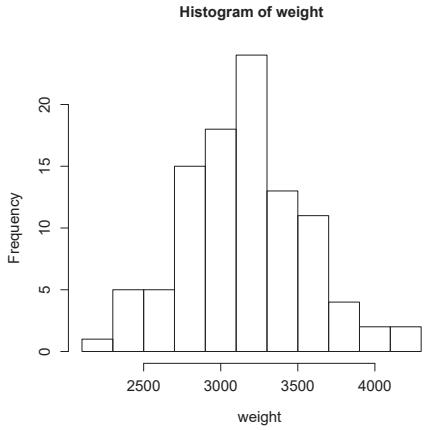

図 9.2　新生児の体重データのヒストグラム (階級全体 $[2100, 4300]$ を階級幅 200 で分割した場合)

さらに，確率密度関数の推定という観点から，ヒストグラムの面積の和が 1 となるように縦軸のスケールを f_j/nh と調整する場合もあり，この場合のヒストグラムの作成にはオプション `freq=FALSE` を与える (図 9.3 を参照)．

```
> hist(weight,breaks=seq(2100,4300,200),freq=FALSE)
```

ヒストグラムはデータの分布を視覚的にとらえるためにはよいが，その形状は階級の設定に依存し，ベストのものを選ぶことは難しい．この問題に対して，データの四分位点の情報を利用して分布を可視化する手法が**ボックスプロット**[8](box plot) である (図 9.4 を参照)．ボックスプロットは，「箱 (ボックス)」と「髭 (ヒゲ)」の部分からなる[9]．ボックス内の太い実線はデータの中央値を表し，ボックスの上端と下端の範囲は，四分位範囲 $Q(= Q_3 - Q_1)$ であり，

8)　ボックスプロットは J. W. Tukey によって提案された (Tukey (1977) を参照)．

9)　この意味で，**箱髭図** (box-and-whisker plot) とよばれることがある．

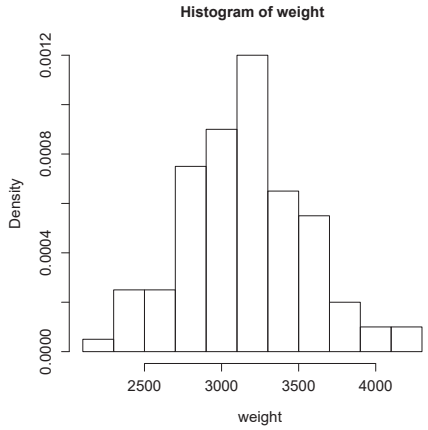

図 9.3 新生児の体重データのヒストグラム (階級全体 $[2100, 4300]$ を階級幅 200 で分割し，縦軸のスケールを面積が 1 となるように調整した場合)

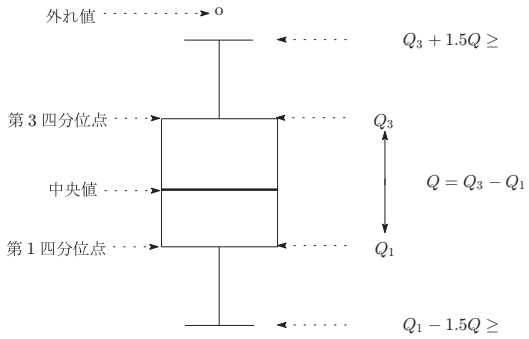

図 9.4 ボックスプロットの一般的な説明

全データのうち中央に位置する 50% のデータが，この範囲に入ることを表している．また髭の範囲は，箱から四分位範囲の幅の 1.5 倍の位置より内側で最も近いデータの位置までを表しており，この範囲外のデータは外れ値とみなされる[10)]．

ヒストグラムは階級を決定する方法が多様に存在するのに対して，ボックスプロットは髭の長さに関する選択の方法によって若干の差異は生じるものの，5 点要約値を使って描くことができ，この意味で，関数 summary から得られる数値的な要約を可視化したものと考えることができる．

R では，関数 boxplot を使ってボックスプロットを描くことができる (図 9.5 を参照)．

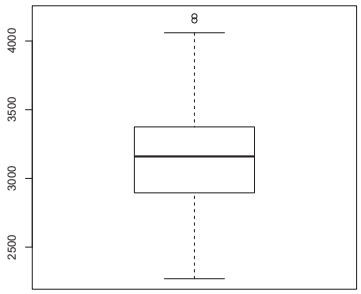

図 9.5 新生児の体重データのボックスプロット

```
> boxplot(weight)
```

この結果から，新生児の体重のデータに関しては，正の外れ値と思われるデータが若干存在

10) 正規分布の場合に「髭の外」にデータがはみ出す確率は約 0.35% であり，「かなり稀」であることから，このようにみなされる．

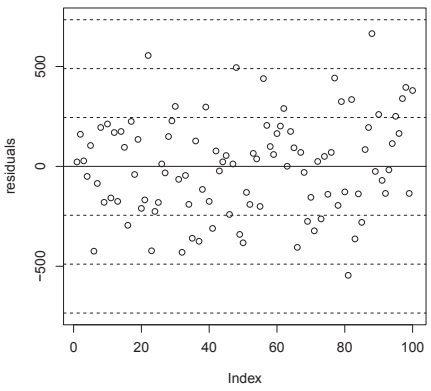

図 13.7　残差のインデックスプロット (単回帰分析の場合)

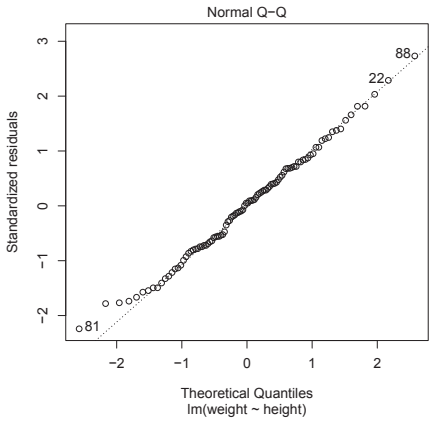

図 13.8　残差の正規 Q-Q プロット (単回帰分析の場合)

から，残差はほぼ直線上に位置し，誤差の正規性を疑う結果は得られないことがわかる．

13.2 重回帰分析　　check box □□□

13.2.1　重回帰モデル

例えば，父親と母親の身長と成人した子どもの身長や，従業員数と資産と売上高，または所得と資産と消費などのように 3 つの変量 (X_1, X_2, Y) を考え，その実現値 (x_1, x_2, y) の間にどのような関係があるかということを考えることは，単回帰分析の自然な拡張である．

ここでは，(x_1, x_2) の線形結合と y の間に線形関係があるかを調べる**重回帰分析** (multiple regression analysis)[23] を扱う．

具体的には，(x_1, x_2) と y との間には以下の関係が成り立つと仮定する．

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon \tag{13.52}$$

これは，**重回帰モデル** (multiple regression model) とよばれ，各項の名称は単回帰モデル (13.1) の場合と同様である．特に，説明変数は x_1 に加えて x_2 が増えており，それにともなって，回帰係数も β_2 が追加されている．

重回帰モデルは，入力を x_1, x_2 とし，出力を y とする (線形) システムという観点からは，図 13.9 のように理解できる．

$$\overset{\text{Input}}{x_1, x_2} \;\rightarrow\; \overset{\text{Linear System}}{\beta_0 + \beta_1 x_1 + \beta_2 x_2} \;\rightarrow\; \overset{\text{Output}}{y}$$
$$\uparrow$$
$$\underset{\epsilon}{\text{Error}}$$

図 13.9　線形システムと重回帰モデル

23)　ここで，「重」は「多重」(複数) の説明変数を使って分析することを表す．

実際に，変数のセット (x_1, x_2, y) に対する n 組の観測値 (3 変量データ) (x_{i1}, x_{i2}, y_i) $(i = 1, \cdots, n)$ が与えられた場合は，重回帰モデル (13.52) は以下のように成分で表される．

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i \qquad (i = 1, \cdots, n) \tag{13.53}$$

この成分表現において，“i” は観測対象である個体番号を表すことに再び留意しよう．

重回帰モデルの成分表現は表 13.2 のようにまとめることができる．

表 13.2 重回帰モデルの成分表現

i	y_i	$=$	β_0	$+$	$\beta_1 x_{i1}$	$+$	$\beta_2 x_{i2}$	$+$	ϵ_i
1	y_1	$=$	β_0	$+$	$\beta_1 x_{11}$	$+$	$\beta_2 x_{12}$	$+$	ϵ_1
2	y_2	$=$	β_0	$+$	$\beta_1 x_{21}$	$+$	$\beta_2 x_{22}$	$+$	ϵ_2
$\vdots$		$\vdots$							
n	y_n	$=$	β_0	$+$	$\beta_1 x_{n1}$	$+$	$\beta_2 x_{n2}$	$+$	ϵ_n

13.2.2 最小自乗法と最小自乗推定値

重回帰モデル (13.53) における回帰係数 β_0, β_1, β_2 は未知母数であるので，単回帰の場合と同様に最小自乗法で推定する．すなわち，誤差平方和

$$\Delta^2(\beta_0, \beta_1, \beta_2) := \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2})^2 \tag{13.54}$$

を β_0, β_1, β_2 に関して最小にするという，以下のような定式化がなされる．

最小自乗法の定式化

$$\Delta^2(\beta_0, \beta_1, \beta_2) \longrightarrow \min_{\beta_0, \beta_1, \beta_2}$$

これも単回帰の場合と同様に，誤差平方和 $\Delta^2(\beta_0, \beta_1, \beta_2)$ の最小化は，極値問題に置き換えて解く方法がとられ，以下のように回帰係数 β_0, β_1, β_2 で誤差平方和を偏微分したものをゼロとおくことによって実行される．

$$\begin{cases} \dfrac{\partial \Delta^2(\beta_0, \beta_1, \beta_2)}{\partial \beta_0} = -2 \displaystyle\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2}) = 0 \\ \dfrac{\partial \Delta^2(\beta_0, \beta_1, \beta_2)}{\partial \beta_1} = -2 \displaystyle\sum_{i=1}^{n} x_{i1} (y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2}) = 0 \\ \dfrac{\partial \Delta^2(\beta_0, \beta_1, \beta_2)}{\partial \beta_2} = -2 \displaystyle\sum_{i=1}^{n} x_{i2} (y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2}) = 0 \end{cases}$$

これらの式を整理することによって，β_0, β_1, β_2 に関する連立 1 次方程式 (正規方程式) が以下のように得られる．

$$\begin{cases} n\beta_0 + \left(\displaystyle\sum_{i=1}^{n} x_{i1} \right) \beta_1 + \left(\displaystyle\sum_{i=1}^{n} x_{i2} \right) \beta_2 = \displaystyle\sum_{i=1}^{n} y_i \\ \left(\displaystyle\sum_{i=1}^{n} x_{i1} \right) \beta_0 + \left(\displaystyle\sum_{i=1}^{n} x_{i1}^2 \right) \beta_1 + \left(\displaystyle\sum_{i=1}^{n} x_{i1} x_{i2} \right) \beta_2 = \displaystyle\sum_{i=1}^{n} x_{i1} y_i \\ \left(\displaystyle\sum_{i=1}^{n} x_{i2} \right) \beta_0 + \left(\displaystyle\sum_{i=1}^{n} x_{i1} x_{i2} \right) \beta_1 + \left(\displaystyle\sum_{i=1}^{n} x_{i2}^2 \right) \beta_2 = \displaystyle\sum_{i=1}^{n} x_{i1} y_i \end{cases} \tag{13.55}$$