まえがき

本書は，初歩のデータ分析からビジネスや研究で必要となる高度なデータ解析まで，読者が的確に行えるようになるための手助けを目指している．したがって，執筆にあたっては厳密さより直観的な理解の助けとなることを重視した．また，自習する読者のため，随所に演習問題を配置した．

本章は 2 部構成となっている．第 I 部は，2012 年度から高等学校の一年次必修科目「数学 I」に単元「データ分析」が導入されたことを踏まえ [18]，その内容に沿ったものとした．最近の高等学校の教科書は驚くほど薄い．これではどうしても要点だけの説明になりがちで，まじめに考える生徒ほどわけがわからなくなってしまうのではないだろうか．特に単元「データ分析」は，その意味や背景まで理解することなしに形式的な理解に終始するだけでは，無味乾燥で退屈な内容としか思われないだろう．したがって，本書の第 I 部は，教師の方々だけでなく，意のある生徒にもぜひ**データの面白さ，奥深さ**を味わっていただきたいと思い執筆した次第である．

新たな事実の発見には的確なデータ解析が欠かせない．他人に対し，それを本当に納得させようとしたときも同じである．さらに，データからわかったことに一般性を持たせるには，モデルの形に昇華することも必要になる．残念ながら，その過程は平坦な一本道ではない．解析できるようにデータを整えることから始まり，可視化によって人間の感性を呼び覚ますことでデータの裏にある現象を探ること，その結果を数式の形のモデルで表現し，他人にうまく伝えることなどさまざまな要素があり，そのいずれにも多くの選択肢がある．

言い換えれば，このような，データから新たな価値を創り出す「データサイエンス」は，さまざまに考えを巡らせ新たな価値を創りだす芸術の側面が強い．英語では科学を Art of Science というが，まさにそれを実践することになる．日本にはこのような芸術家肌の科学者が圧倒的に不足している．本書が，多くの方が実践に取り組んでいただくきっかけとなり，数多くの先進的なデータサイエンティストが育つことにつながれば，望外の幸せである．

第 II 部はデータ分析からデータサイエンスへの橋渡しとなることを目指している．慶應義塾大学理工学部で実施してきた演習つきの講義科目「統計科学同演

習」の内容をベースとし，大学院総合教育科目「データリテラシー」や学部科目「生命科学のための確率論」の講義内容，さらには早稲田大学理工学術院での講義内容も適宜取り入れた．そのためデータサイエンスの最終目的である，新たな価値の創造まで十分踏み込めてはいないが，データサイエンスの一つの入口になるに違いない．

この第 II 部は著者が 20 年来提唱してきた「データサイエンスの入門と実践」でもある．まず第 6 章でデータサイエンスを概説する．ここを一通り理解していただければ，「データサイエンティスト」への道が開けるものと信じている．さらに第 7 章以降では．さまざまな分野にわたる実データの解析を紹介している．つまりデータサイエンスの実践例である．したがって，興味のある章から始め，必要に応じて他の章を参照しながら，ぜひ実際に手を動かしてみていただきたい．

また，2013 年に設立した（株)データサイエンスコンソーシアムの事業展開に伴い，お付き合いさせていただいた数々の日本を代表する企業から，企業におけるデータサイエンス実践の現状もいろいろ教えていただいた．この貴重な経験を踏まえ，本書，とりわけ第 II 部が各企業でのスムーズなデータサイエンス実践の助けにもなるよう十分配慮した．

この内容をマスターすれば，データサイエンスの基礎は身に付けることができるに違いない．つまり，実データを解析することのむずかしさや厄介さを理解し，それを乗り越えるために必要なスキルを身に着けた，底力のあるデータサイエンティスト誕生の助けとなるだろう．「生兵法は大怪我のもと」は頭に留めておいてほしい．

データは現象の放つ光である.

現象の放つ光をどれだけうまくとらえ，結果に結び付けるか，それがデータ解析の神髄であることは，本書を読み進むうちにきっとおわかりいただけるに違いない．

丸い囲み記事 はコメント，四角い囲み記事 はトピックの解説，影つきの四角い囲み記事はその節の簡単なまとめであるので，最初は読み飛ばしても差し支えない．読み進んでいくうち，あるいは読み終えたあとで，必要に応じて参照していただければ幸いである．

なお実際にデータを扱うにはソフトウェアを必要とする．もちろん Excel など表計算ソフトでも，工夫次第である程度のことまではできるが，より柔軟にデータを探索するにはそれなりのソフトウェアを用いることをおすすめする．ここでは，探索的データ解析を提唱した J.W. Tukey [50] の弟子たちが，ベル研究所を拠点に柔軟なデータ解析環境を一から作り直したソフトウェア S [6, 43] のパブリックドメイン版である R を基本的なソフトウェアとして用いる．

この R は，だれでも自由に (http://www.r-project.org) あるいはそのミラーサイトからダウンロードし利用できる．R に関してはさまざまなレベルの入門書や RWiki など Web アクセスできるサイトも存在するので，使い方に習熟するのには特に問題はないだろう．

関連する，R の関数，データなどのオブジェクト名あるいはファイル名を，欄外にタイプライタ文字で記した．関数の場合は名前の後にカッコが続き，必要に応じて追加引数がその中に置かれている．また，波カッコが続いていれば，その名前のライブラリをインストールする必要があることを示している．ただし DSC はデータサイエンスコンソーシアム (http://datascience.jp) から入手できる R ダンプファイルに含まれる，R オブジェクトや Zip ファイルである．区別のため，DSC から入手したオブジェクトには，すべて，大文字ではじまる名前が付いている．

ライブラリーのインストールは R のメニューバーのパッケージで「パッケージのインストール」を選択した後，ミラーサイトを選択し，ライブラリー名のパッケージを指定すればよい．インストールしたライブラリーを用いるには関数 library でライブラリーを指定する必要がある．一方，ファイル名に拡張子 R が付いた R ダンプファイルの場合は関数 source で読み込むだけで，すぐ使えるようになる．なお，本文中の > で始まる式は, そこに示された図などを描くための R 式である．

索引での太字の数字は，参照頁のうちでも基本的な説明のある頁を示している．本文中では，このような用語は，太字になっているだけでなく，カッコ書きで英語も付されているので区別は容易である．ぜひ活用していただきたい．

執筆にあたっては読みやすさを重視し，『科学技術系のライティング技法』[23] の著書もある 近代科学社の小山透氏に厳しくチェックしていただいた．また，慶應義塾湘南藤沢中高等部の馬場国博君，一橋大学大学院国際企業戦略研究科の横内大介君，St Andrews 大学の島津秀康君，村田事務所の村田恵君，(株) データサイエンスコンソーシアムの仲真弓君，早稲田大学国際教養学部の力丸佑紀君，東京大学経済学研究科の菅澤翔之助君，早稲田大学理工学術院の飯島泰平君にも校閲をしていただいた．また，近代科学社の小山透氏と高山哲司氏には，出版にこぎつけるまで辛抱強くご支援いただいた．

本書がすこしでも読みやすくなっているとすれば，それは，いずれも上記諸氏のご尽力の賜物である．ここに深く感謝する．

2015 年 10 月
湘南国際村にて

カバーデザイン

本書のカバーには，著者の両親と長年親交のあった画家，中神潔氏が 1991 年に描かれた『いまきたこの道』と題する油彩を使わせていただいた．中神潔氏が，富士山麓の別荘地に建てた小屋を拠点として，スケッチを重ねた作品の一つがこれである．小屋の脇の急な坂を上り切ったところで，いまきた道を振り返っている少女の目にとまったのは小さなタンポポの花であろうか．そんなメルヘンが氏の作品の特徴である．

情熱を込めたものでなければ感動を生まない．

一つの作品が出来上がるまでには長い道のりがある．すこしずつ絵具を重ねるだけでなく，納得いかなければ描き直す．その繰り返しで，ようやっと一つの作品が完成する．それを支えるのは，確実な技能と尽きることのない情熱であろう．

裏カバーには，同じ中神潔氏の 1983 年の作品『富士夕景』を使わせていただいた．いまや，さまざまなツールをうまく使いこなせば，それなりの作品を作り出すこともできる時代である．しかし，そんな道具に頼り切らず，ぜひ高みを目指してほしい．

著者の似顔絵

奥付に使わさせていただいた著者の似顔絵は，慶應義塾大学理工学部で 20 年以上の歴史を持つデータサイエンス研究室の卒業生で，研究室の秘書も長年勤めていただいた宮澤美紀君の作である．

目 次

第 I 部　データ分析

第I部

データ分析

第 I 部を執筆するにあたっては,「データリテラシー」[39], つまりデータを分析するときに最低限心得ておく必要のある常識を, スムーズに習得できるよう留意した. データは現象の放つ光であり, さまざまな色合いで輝いている. 決して無味乾燥なものではない. いかにうまくその輝きをとらえるかがデータ分析の価値を大きく左右する. 第 I 部を読み進むことでデータをより深く理解できるようになれば, それは成功への第一歩である.

データに含まれるランダム性を考慮することも重要であるが, それは第 II 部に譲り, ここではもっぱら, データ, 特に大量なデータを扱うときに心得ておくべき基本的な事柄を習得していただくことを主眼にする. そのため, あまり具体的な問題やデータをそのままの形で取り上げることはしない. 内容がいくぶん単調に感じられたら, 第 I 部を途中でスキップして第 II 部を覗いてみるのもよい. その手がかりを随所にコメントとして付加することに努めた. 第 I 部の内容がまた違って見えてくるに違いない.

第 I 部は高等学校の一年次必修科目「数学 I」の最後の単元「データ分析」の内容に沿って話を進めるので, それまでの「数学 I」の各単元の内容と連携することにも留意した. たとえば, この単元に至るまでに, 実数, 1 次不等式, 集合と論理, 関数とグラフ, 2 次関数などさまざまな単元をすでに習得しているはずであるが, 必ずしもそれが血となり肉となるところまでには至っていないかもしれない. この第 I 部では, それらを総動員して話を進める. ただし, 高校生の範囲を超えている部分もあるので, そのような箇所は適宜読み飛ばしていただきたい.

第1章　データ

ちかごろ「データにもとづく客観的な判断が必要」と耳にすることが多い．また「データサイエンス」,「データサイエンティスト」などという言葉がマスコミにしばしば登場している．[1)]　そのため，データ主導型 (data-driven) の経営やマーケティング，意思決定が重要であると強調されることも多い．最近では，安価なコンピュータを多数連結することで，きわめて大量なデータの蓄積や処理が可能になってきたこともそれを後押ししている．使い道がはっきりしていないデータでもとりあえず貯えておけば何らかの形で役立つだろうという期待から，いわゆる**ビッグデータ** [2)] の取扱い技術も注目されている．

しかし，同じ「データ」といってもそれが具体的に何を指し示すのか，場合により人により大きく異なり，混乱の元になっていることが多い．そこで，まず，本書で「データ」といったとき，何を指し示すのかはっきりさせておくことにしよう．たとえば『新明解国語辞典（第 7 版)』(以下『新明解』と略記）[42] では，「データ」は

1. 推論の基礎となる事実
2. その事柄に関する個々の事実を，記号（数字，文字，符号，音声など）で表現したもの （もっとも狭い意味では，数値で表現したものを指し，広義では参考となる資料や記事のこともいう．また，コンピュータの分野では，処理できる対象すべてを指す．したがってプログラム自体もデータであるが狭義では除外する）

となっている．

第 1 義は「データを示す」あるいは「万全のデータがそろった」といった使い方の場合で,「根拠を示す」あるいは「完璧な根拠がある」ことを意味している．第 2 義はより具体的に，推論の根拠となる報告書などの文書あるいはファイルを「データ」としており,「推論の根拠となる資料 (material)」の言換えである．音声データ，画像データなども当然ここに含まれる．いずれにせよ，この辞典の説明のポイントは，**データは推論の根拠となるものである**と強調している点にある．

1) 第 6 章が密接に関係する．

2) 典型的な規模の大きなデータとしては，衛星から連続的に送られてくる各種センサの示す値や画像・レーダーデータ，ゲノムデータ，原子炉などの複雑な実験データ，インターネットを流れるパケットデータ，電力のスマートグリッドデータなどがあるが，英語の big には「規模の大きな」という意味だけでなく「扱いの大変な」あるいは「重要な」という意味もあり，ビッグデータはこの二つを合わせて意味している．

何でもデータ？

『新明解』[42] の第 2 義にあるように，コンピュータサイエンスの分野では，コンピュータで扱えるものは何でもデータとみなすという割り切りが行われ，対象の抽象化に大いに役立った．「データ駆動型プログラミング (data-driven programming)」はその成功例である．しかし，このことによってデータの重要な側面である「推論の根拠」が忘れ去られ，データは何ともとりとめのないものになってしまった．データという言葉の乱用のもたらした副作用である．

本書では守備範囲を明確にするため，第 2 義の「数値，記号で表した推論の根拠」を広義のデータと呼ぶことにする．音声でも画像でも結局，数値あるいは記号で表せるので，一般性は失わない．しかし，これでも扱う対象が数値や記号に絞られただけで，分析の対象としてつかみどころがないことに変わりがない．もう少し，何に注目してどう推論するのか枠組みをはっきりさせた上でのデータ分析でない限り，的確な分析は望むべくもないし，新たな価値を生みだすこともかなわない．特に大量で複雑なデータの場合はそうである．

データの構成を明確にするのに **変量** (variate) の概念が役立つ．たとえば，ある地区に登録されている車のデータを考えてみよう．この場合，車検証にある登録番号，登録年月日，初年度登録年月日，種別，自家用・事業用の別，...，使用の本拠の位置，有効期間の満了する日，備考の項目が変量と考えられる．さまざまな車について，車検証に記載されている変量の値を集めたものが，その地区に登録されている車のデータである．

このように，変量を導入すれば，対象とするデータがどのような値からなる記録かを簡潔に表現できるだけでなく，どの変量とどの変量に注目して分析すればよいのか，その方向性を明確にするのにも役立つ．これらの変量のなかには，「型式」と「長さ」，「幅」，「高さ」のように，ある変量の値が他の変量の値を定めてしまう，いわば主従関係にある変量や，「登録番号」のように記録全体を一意に定めてしまう **ID** の役割を果たす変量もあり，その性格はさまざまである．さらに，「備考」のように値の形式が定まっていない変量すら存在する．このような各変量の特徴をよく理解することが分析の方向性を定めるのに大いに役立つ．

一般的にどのような変量を導入したらよいかは，現象とそこから得られるデータがどのようなものか，どのような目的でデータ分析を行いたいかなどに大きく依存し一定ではない．また解析する人間の考え方も反映される．したがって，ごく単純な場合は別として，一般的にはこの変量の導入の段階で大いに頭を悩まし時間を費やすことになる．しかし，それで得るものは大きく，現象や問題の見方

特に，$a = 1/s(x)$，$b = -\bar{x}/s(x)$ に取れば，いつでも $\bar{y} = 0$，$s(y) = 1$ となるように線形変換できる．つまり，

$$y_i = \frac{x_i - \bar{x}}{s(x)},\ i = 1, 2, \ldots, n$$

によって，常にデータ y の平均値を 0，標準偏差を 1 に揃えられる．これは，古くから**データの規準化** (data normalization) として知られている方法であるが，偏差値はこれを積極的に利用して，テストの難易度やクラスの出来の違いを平均と分散という観点から取り除いている．

具体的には 100 点満点のテストを想定して，平均値を 50 に，標準偏差を 10 にそろえる規準化

$$y_i = 10\,\frac{x_i - \bar{x}}{s(x)} + 50,\ i = 1, 2, \ldots, n$$

をしたものが偏差値である．[39)]

問題 1.4.1 偏差値データ y の平均が 50, 標準偏差が 10 であることを示しなさい．

39) 原理的には，偏差値は 100 を超えることも，負になることもありうる．

偏差値の功罪

このような規準化によって失うものも大きい．たとえば $\bar{x}$，$s(x)$ の値が保存されていなければ，データ y をデータ x に戻すことはできない．つまり，テストのあと，偏差値だけしか知らされなかったら，自分がそのテストでどれだけの得点をし，どれだけ実力を発揮できたのかわからず，何とも味気ないことになる．それだけでなく，クラスの皆がよく出来たらほめるという本来の教育の姿に反して，皆が似たような成績だと，ほんの 1,2 点の差が偏差値には大きく反映してしまい，つまらない競争心をあおる原因になりかねない．また，科目ごとの偏差値を足し合わせて総合偏差値とすることもよくあるが，これは各科目にいつも標準偏差の逆数のウエイトをつけて総合することになる．その結果，標準偏差の大きな科目ほど総合偏差値に占める割合は小さくなり，標準偏差の大きな科目，つまり感度のよいテスト [40)] ほど，過小評価されるといったことが起きてしまう．最近では，テストの成績に限らず何でもまず偏差値に直してからという変な習慣が見受けられるが，そこにはさまざまな危険が潜んでいることをよく認識する必要がある．6.10 節ではこのような規準化をさらに一般的に議論する．

40) ここでは，生徒の学力の差をどれだけ敏感に反映するテストであったかをテストの感度と呼んでいる．

線形変換は，$a > 0$ である限り，データ x_i, $i - 1, 2, \ldots, n$, の相対的な位置関係を変えないので，相対的な関係に重きがあるときには有効な変換である．テス

が開始されていたことも読み取れる．

箱ひげ図が頻繁に使われるようになったのは，Tukey [50] による**探索的データ解析** (EDA) の提唱をきっかけにしてであるが，ロウソク足はそれよりずっと以前の江戸時代から米相場の視覚表現として用いられてきた． 箱とひげの組合せでデータをわかりやすく表現するという先駆的なアイディアが，他ならぬ日本で生まれたのは誇ってよい事実である．この背景には，大阪・堂島の米市場が先物取引を世界に先駆けて開始していたほど成熟度の高い市場であったことが大きい．図 4.7 のような日本ではなじみ深い，駅やバス停の，各時刻ごとに何分に電

図 4.7 バスの時刻表

車やバスが来るかがわかる時刻表も，電車やバスの到着時刻の概略と詳細の両方を伝えられる優れた視覚表現であるが，Tukey はこれと同じアイディアの図 4.8 のような表示を**幹葉表示** (stem and leaf display) と称し，世界中に広めた．縦棒の左の幹に相当する部分にはデータの値の先頭桁の数字が置かれ，縦棒の右の葉に相当する部分には次の一桁の数字が並んでいる．幹に同じ数字が重複しているのは，葉の数字が 4 以下の場合と 5 以上の場合に分けて表示しているからである．

`stem()`

図 4.8 は 3.2 節で扱った『音楽アルバムの週間売り上げ数トップ 300』データの幹葉表示であるが，表中の `The decimal point is 4 digit(s) to the right of the |` は，小数点が縦棒の 4 桁右にあることを示している．つまり棒の左の数字の右に小数点を置き，その右に棒の右の数字一つを置いてできる小数を

```
The decimal point is 4 digit(s) to the right of the |

  0 | 0000000000000000000000000000000000000000000000000000000000000000000000000000+193
  0 | 55556666667788889
  1 | 003
  1 | 56
  2 | 4
  2 |
  3 | 0
  3 | 5
  4 |
  4 |
  5 |
  5 | 77
```

図 4.8 幹葉表示

10^4 倍した値が，2 桁の精度で切り捨てたとき該当するデータの値一つを表している．たとえば，最後の行の `5 | 77` は，5.7×10^4 以上 5.8×10^4 未満の枚数，つまり 57,000 枚から 57,999 枚の間の売上枚数のアルバムが 2 アルバムあることを示している．

また，最初の行には棒の右に 0 が 80 個並んでいるが，これは 0.0×10^4 以上 0.1×10^4 未満の枚数，つまり 0 枚から 1,000 枚未満売り上げのアルバムがすくなくとも 80 アルバムあることを示している．この行の右の`+193` は，葉として並べきれないアルバムが 193 アルバムあることを示している.[10] 省略されているのは 0 から 4 までの数字であるので，5,000 枚未満売上のアルバム 193 件が表示省略されていることになる．

[10] 関数 `stem` で 1 行に表示できる葉の最大数はディフォルトで 80 になっている．

同じデータの箱ひげ図である図 4.4 と比較してみればわかるように，全体的にはヒストグラムあるいは棒グラフの役割を果たし，より詳しい値も調べられるという点で，単純でありながら優れた視覚表現である．しかし歴史的には日本の時刻表のほうがずっと古い．Tukey の弟子であるベル研究所の研究者 Rick Becker (図 4.9) が著者を訪ねて来日したとき，日本の時刻表をみて大きなショックを受けていたことが思い出される．

日本人のセンス，世界に冠たるものである．データ解析用の計算言語 R や，S-PLUS の元である S の開発は彼らによるものであるが，その開発過程で，ある日本人研究者の大きな寄与があったのも偶然ではない．

図 4.9 S 開発の主役のひとり Rick Becker と，彼らを助けた日本人研究者

データ分布の視覚表現

箱ひげ図やロウソク足に代表されるデータ分布の視覚表現は，箱やひげといった単純な部品で必要な情報を表現し，人間の直観に訴えるところにある．したがって，どのような視覚表現が適切かは，データの性格だけでなく，その使用目的にも大きく依存する．

4.3 ひとこと

第 10 章で紹介するような，株価を代表とする金融データの解析やファイナンス理論の実装にあたっては，**差分** (difference) が重要な役割を果たす．そこで，ここで差分について，すこし説明しておくことにする．第 I 部の範囲を多少超えるところもあるので，読み飛ばしていただいても構わない．

数列 $x_1, x_2, \ldots, x_n$ の差分[11] には，**前進差分** (forward difference) $y_k = x_{k+1} - x_k, k = 1, 2, \ldots, n-1,$ と**後進差分** (backward difference) $y_k = x_k - x_{k-1}, k = 2, \ldots, n,$ がある．どちらでも得られる系列は同一であるが，k が時刻などに対応している場合は，前進差分と後進差分では k の扱いが異なる．前進差分の場合は最後の時刻を除き，後進差分の場合は最初の時刻を除くことになる．差分を取ると，ゆっくりした動きは除かれる．たとえば，数列 $\{x_k = ak + b,\ k = 1, 2, \ldots, n\}$ の差分は定数 a となるので，数列 $x_1, x_2, \ldots, x_n$ に含まれる直線的な動きは動きのない定数となり，除かれる．

[11] 差分をあらわすのに記号 Δx_k が用いられるが，前進差分のこともあれば後進差分のこともある．

`diff()`

また，差分は導関数を数値的に近似する一つの手段でもある．たとえば $0 \leq t \leq 1$ 上の関数 $x(t)$ の導関数 $x'(t)$ は，数列 $\{t_k = k/n, k = 1, 2, \ldots, n\}$ と

第II部

データサイエンス

第 II 部はデータから新たな価値を生み出すデータサイエンスの基本を理解していただくことを主要な目的とする．データの背景に潜む現象をうまく浮かび上がらせ，データの価値を最大限引き出すためには，第 I 部で身に着けたデータリテラシーが大いに役立つ．また，解析しやすいようデータを整えたり，データの概略をつかみ解析の方向性を見極めるにも，データリテラシーの高さがものをいう．

第 6 章はデータサイエンス入門である．データサイエンスとは何かから始め，その実践にあたっては，どのような知識や経験が必要になるか，さまざまな側面から解説する．第 7 章は個体の雲の探索である．データの雲も個体空間で眺めるか変量空間で眺めるかによって見えるものが違ってくる．ここでは，主に個体空間で眺める．このような探索によって，データの様子がある程度つかめてくれば，次の解析のストラテジーもはっきりさせられる．第 8 章は変量間の関係である．変量の間の関係をつかむことが，より的確なモデリングに結びつく．

そして，第 9 章は変量間の相関である．相関は，関係のようなはっきりしたものではなく，一種の傾向，つまり関連性でしかないが，見当をつけるには大いに役立つ．最後の第 10 章は確率モデルに進む．まったく異なるデータであっても，同じようなアプローチでその裏に潜む現象をモデリングできるのが，確率モデルの特徴である．地震データと株価データを中心に確率モデルの理解を深める．

紙幅の関係で，ここではごく初歩的なモデルしか紹介できなかった．データの多様性を反映しモデルも多様である．また，日々新しいモデルが生み出されている．それらを整理した形でのデータサイエンスの実践は別の機会にゆずりたい．

モデルは数式による表現である「数理モデル」の形をとることが多く，線形代数，微分積分，確率論などの基本的な数学の利用は避けられない．データを解析することで何かを明らかにしたいと思っていても，数学は苦手で敬遠したいと思っている方も多いのではないだろうか．

そこで，本書では第 9 章の前半までは線形代数の知識だけで理解できるよう配慮した．もちろん，線形代数を使わなくても説明は可能であるが，そのことによる煩雑さ，見通しの悪さを考えれば，線形代数を使うことのメリットは大きい．たとえば，基底の概念や射影の概念を理解してしまえば，最小二乗法の理解も容易となる．しかし，それを避けることによって，あまり本質でない計算に煩わされ，直観的な理解にまで至らない恐れのほうが大きい．直観的な理解なしで，複雑で大規模なデータ，いわゆるビッグデータに取り組んでも途方に暮れるだけであろう．その意味でも，この機会に線形代数にも慣れ親しんでいただければと念じている．

この第 II 部で展開される様々な実際問題とそのデータ解析，そこに現れるモデルの楽しさを実感していただくことで，数学の面白さ，強力さも再認識していただくきっかけをつかんでいただければ，これに勝る幸せはない．数学は厳密さも重要であるが，「数式から問題の本質をつかめるようになること」，「自分が創り上げたモデルを数式で表現できるようになること」のほうがもっと重要である．

本書は，第 I 部を前提知識として，大学あるいは大学院の教科書としてして使うこともできるように構成されている．すでに確率論を習得している学生が対象ならば，最後の第 10 章から始めランダムな現象のモデルの実際を理解させたあと第 6 章に戻り，実際のデータ解析が確率論の枠組みだけで片付くわけではないことを，学生に理解させることができれば，真のデータサイエンティスト育成コースとなることであろう．その結果，第 7 章から第 9 章までの内容も，より深く理解できるようになるに違いない．

社会人の方が自習するときには，まず第 7 章までを読み進み，必要に応じて残りの章を読むとよい．実データを用いた例や演習問題が理解を深めるのに役立てば幸いである．

第6章　データサイエンス入門

第 I 部は「データ分析」と題して話を進めてきた．これは，高等学校の「数学 I」の単元名にならったからであるが，再び『新明解』[42] を参照してみれば，「分析」は「単純な要素にいったん分解し，全体の構成の究明に役立てること」とある．これからすると，第 I 部は，データをいくつかの単純な要素に分解する「データ分析」の初歩的な道具である代表値や箱ひげ図，散布図と相関係数などを，断片的に紹介したにすぎない．第 II 部は，これを更に発展させ，効果的な**データ解析** (data analysis) によって新たな価値を生み出す，**データサイエンスの実践**を目標とする．[1)]

[1)] データサイエンスあるいはデータサイエンティストという言葉で，具体的に何を表しているかについては，6.4 節，6.5 節を参照されたい．

解析

数学の研究分野の一つに**解析学** (analysis) がある．この名前の由来は，一つの関数を無限個の項の和に分解し，研究するところにある．もちろん，解析学は単に関数を分解するだけでおしまいではない．その無限個の項のうち，どの項まで注目すれば全体を説明するには十分か，また，そのために必要な条件は何かを探り，それらをまとめあげることで，その関数の性質を明らかにしようとする．データ解析も同様で，データをいったん単純な要素に分解し，どの要素が，その背後にある**現象** (phenomenon)[2)] のどの部分をどれだけ説明できるか探り，それらをまとめあげることで，全体像を明らかにする．したがって，データ分析とデータ解析の違いは，簡単に言えば，要素に分解したままで済ませるか，さらに現象の解明にとって重要な要素だけを抜き出し，まとめ上げるところまで進むかどうか，であるといってもよい．

[2)] ここでの**現象**は，「起きていること」の総称である．物理現象，化学現象などに限らず，製品の製造，販売，病気，犯罪など社会現象，経済現象など，すべての現象を含む．

まず，本章では，データ解析を効果的に行うためには欠かせない **データサイエンスの枠組み** (framework of data science) と，その背景を説明する．なお，以下では，わかりやすさを優先して，数学としては欠かせない細かな条件を，必ずしも明記していないので，必要に応じて，巻末に掲げた専門書 [16,17,21,37,47] などを参照するなどして補足していただきたい．

6.1 データに語らせる

データから何かを引き出すには，データに「語らせる」ことが重要である．データを**現象の放つ光**と思って，じっくり眺めているうちにデータは語り始める．眺めるといっても，「こだわり」を捨てなければ本当の姿は見えてこない．こだわりの中には，特定の理論や特定の手法などに対するこだわりも含まれる．データと対峙することで得られる最大の収穫は，自分の思っていたことが，果たして本当か，単なる思い込みだったのかが判明することであろう．このときこそが，**霧が晴れたという感動**を味わえる貴重な一瞬である．山登りと同じで，データサイエンティストは，この瞬間が忘れられなくて，また新たなデータと格闘し始めるのかもしれない．**データと真摯に向き合う**ことが，データからの新たな発見や価値の創造への近道であることは，ぜひ心に留めておいてほしい．[3)]

現象の放つ光を注意深く眺めれば，自然と，それをどう料理したらよいか見えてくる．レシピにしたがって，決められた手順で調理するのも一つの料理ではあるが，それではプロの料理にはならない．素材の味を生かす創意工夫があってこそプロの料理と言えるに違いない．素材の味を生かす料理は「データの背後にある現象をうまく浮かび上がらせる工夫」に相当する．

とはいえ，料理にしてもデータ解析にしても，まずは決められたレシピをこなすことから始め，次第に腕を上げていくしかない．第 II 部は，そのようなレシピのいくつかを紹介できたらという思いで執筆した．あとは，読者諸兄姉が精進して腕を上げ，皆を感動させるようなデータ解析ができるようになれば本望である．ぜひ，**データ解析のプロ**をめざしてほしい．

プロは誇りをもって仕事をする．**データサイエンティスト**もさすがプロといわれるような仕事をしてほしい．[4)] プロの料理人は，包丁，なべ，オーブンなど道具の使い方を熟知した上で，肉，野菜，魚，米などの材料を吟味し，砂糖，塩，酢，醤油，こしょうなどの調味料をうまく使い，お客をうならせる料理を出す．**データに語らせる**のも同じで，食材がデータに変わるだけである．

6.2 ストラテジー

データの背後に潜む現象を何もかも明らかにしようとしたら，いくら時間と費用があっても足りない．いつも，どのような切り口で解析するのか，どこまで掘り下げるのか，といった**ストラテジー**（戦略，strategy）を**練り直しながら**進める必要がある．特に**ビッグデータ** (big data) [5)] の場合には，周到かつ大胆なストラテジーが欠かせない．さもないとデータの海に溺れてしまいかねない．塩野七生氏の著書 [45] の一文を引用させていただけば，

3) データと対峙する過程で「データに裏切られた」と思いたくなることがしばしばある．データに照らし合わせたら，当初の見込みに対して否定的な結論しか出てこない．こんなときデータに裏切られたと思いたくなるのも人情ではある．もちろんデータが裏切ったわけではない．見込み違いだっただけである．

4) データはだれでもその気になりさえすれば解析してみることができる．これはちょうど，包丁やなべさえあれば誰でも料理を始められるのに似ている．電子レンジだけでも料理できる時代になったからこそ，プロの料理人の腕が光ってくる．

5) 英語の big には，規模が大きいだけでなく，重大な，重要な，大変なという意味も含まれている．このように理解しないと 2012 年 3 月の米国のオバマ大統領の教書「Big Data Initiative」の真意も理解できないであろう．

> 情報とは，量が多ければそれをもとにして下す判断もより正確度が増す，とは，まったくの誤解である．情報は，たとえ与えられる量がすくなくても，その意味を素早く正確に読み取る能力を持った人の手に渡ったときに，初めて活きる．

しかし，知らず知らずのうちに蓄積されたデータの場合には，そこから何がわかるのかさえ見当がつかないことも多い．そのような場合には，まずデータがどこにどのような形で存在し，管理されているのかをよく把握した上で，ていねいなデータの**ブラウジング** (browsing) [6] を重ね，ストラテジーを練りあげるしかない．

このブラウジング自体，戦略的に行う必要がある．特に，対象とするデータが大規模で複雑に絡み合っている場合には，その絡み合いを解きほぐしながら様子を探っていかねばならず，利用できる記録すべてをブラウジングの対象にしたら途方に暮れるのが落ちである．

すでに第 1 章で述べたように，適切な変量を導入し，データをうまく組み立て上げることが，絡み合いを解きほぐすための第一歩であるが，その結果，いくら記録数が多くても，扱いやすい大きさの記録に**サンプリング**すれば十分であることが判明することも多い．そうなれば，いくら**ビッグデータ**であっても，その記録数の多さに時間と手間をとられずに，肝心の解析をより一層深めることができる．6.8.1 節にサンプリングの例を掲げておいたが，サンプリングの効用の詳細は別の機会にゆずりたい．

どのような場合でも，効果的なストラテジーの確立には，具体的な目標設定が欠かせない．6.4 節で述べるように，データサイエンス実践の目的は，データからの新たな価値の創造，いわば宝を見つけ出すことであるが，何が宝なのかはっきりさせないと，見当はずれのストラテジーを立てることになってしまう．たとえば，砂漠では宝石よりも水のほうが貴重であるし，石炭が黒ダイヤと呼ばれ貴重視された時代もあった．価値は時代により状況により変化する．したがって，何がわかれば新たな価値といえるのか，それをどう利用するのか，といったことまで見通した上で，よくストラテジーを検討する必要がある．[7]

特に，データの背後にある現象がどんなものか分かればよい場合と，将来まで見通したい場合とでは，その目標設定は大きく異なる．前者ならば，次々と浮かびあがる「なぜ」に何らかの答えを見出していけばよいが，後者の場合は，いわば，データという「過去から現在に至るまでの光」から，「将来それがどのような光を放つことになるのかを予測し評価する」ことであるので，単に「なぜ」に答えを出していくだけではすまない．まず，その予測の評価基準を明確にした上で，それに合ったストラテジーを立てる必要がある．

[6] ブラウジングは牛などが草を食べ歩く様子を示す言葉である．転じて「本などにざっと目を通す」という意味でも使われている．ブラウザ (browser) はこの動作を助けるものであり，今日ではウエブブラウザ (web browser) としてなじみ深い言葉になっている．

[7] かといって，検討を重ねすぎるのも考え物である．ある程度の検討が済んだら，直観や想像力を信じ，先に進むのがよい．ストラテジーはいつでも練り直せばよいのである．

おおまかなストラテジーさえ確立できれば，データの整合性をチェックし，追加データが必要ではないかなどを調べ，不必要なデータを削り，解析しやすい形にデータを整える，いわゆる**クリーニング** (cleaning) あるいは **クレンジング** (cleansing) とよばれる作業に入る．この作業は，実際にデータ解析に携わったことのある方なら誰でもが痛感するように，実に手間のかかる作業である．データの取得から新たな価値の創出に至るまでの時間の，半分以上がこの作業に費やされることも珍しくない．よく教科書で例題として取り上げられるような，**きれいなデータ** (organized data) だけを扱えばすむのは，むしろ稀で，誤り，不整合，欠損，不定形，冗長，不明確など，ありとあらゆる問題が潜んでいる**きたないデータ** (disorganized data) を扱わなければならないのが普通である.[8)]

8) といいながら，本書でも例題として取り上げられるのは，わかりやすさを優先し，ある程度きれいなデータに限っている．

クリーニングの作業では，地道に一つずつ解決していくしかないが，システマティックに，そのような経験を蓄積することで，この作業をすこしでも助け，時間を短縮することができないかという思いで始めた研究である DandD については，6.8 節で紹介する．

いずれにせよ，ブラウジングや解析の初期段階では，こだわりを捨て，データを様々な角度から，**無心に眺める**ことが肝心である．そのことで，「解析者の直観」が研ぎ澄まされてくる．その結果を形にするのが，次節で述べる「モデル」である．モデルは，先に述べたように，データを眺める**レンズ**の役割も果たす．うまいモデルさえ導入できれば，それまでおぼろげであったことが, より鮮明に浮かび上がってくる．

ストラテジー

データサイエンスの実践にあたっては，いつも，どのような切り口で解析するのか，どこまで掘り下げるのか，といったストラテジーを，よく練り直しながら進める必要がある．さもないと，データの海を漂い，溺れてしまうことになりかねない．適切なストラテジーの確立には，具体的な目標設定をおこなった上で，さまざまな角度から，データを無心に眺め廻すことが欠かせない．

6.3 モデル

モデル (model) の日本語は「模型」である.[9)] 本物が大きすぎたり，高価だったりして，そのまま扱うのが困難なとき，その代わりの役目を果たすのが模型である．模型というと，プラモデルや模型飛行機，建築模型などを思い浮かべる方も多いかもしれないが，たとえば「地図」も立派な模型である．街を実際に歩い

9) 第 I 部で紹介したさまざまなデータの代表値は，モデルを構成する基本要素ではあるが，モデルは代表値だけでは表せない部分まで表現する．建物でも，面積からだけでは，それが住宅なのかオフィスなのかまではわからないが，模型をみれば一目瞭然である．

て説明するとしたら大変なことも，地図を一枚広げて説明すれば済む．

しかし，同じ街の地図といってもさまざまである．国土地理院の地図と道路地図や観光案内用の地図では，その様子がずいぶん異なる．国土地理院の地図は方角や大きさが正確でなければならないが，道路地図や観光案内地図は正確さより，ドライバーや観光客にわかりやすいほうが優先される．

これから紹介するモデルも基本的には同じである．データの背後にひそむ現象は，そのままでは複雑すぎたり，調べ尽くすには莫大な時間と費用がかかることが多い．そんなとき，代わりの役目を果たすモデルつまり模型が役立つ．もちろん，地図の例を思い浮かべていただけばわかるように，どのようなモデルがよいかは，その目的によって変化する．特に，モデルを創り出す段階で，ぜひ心に留めていただきたいことは，「必要以上に複雑なモデルを作らないほうがよい」というケチの原理 (principle of parsimony) である．

ケチの原理

ケチの原理はオッカムの剃刀 (Ockham's razor) とも呼ばれるもので，「ひとつのモデルで何もかも説明しようとするな．単純なモデルで説明できればそれに越したことはない」という一種の精神論である．たしかに複雑でファンシーなモデルは魅力的ではあるが，表現できた気がするだけで，あまり意味のない結果に終わる可能性のほうが大きい．特に，限られたデータから帰納するときには，そのデータの含む情報量に見合った複雑さのモデルである必要がある．

もともとスコラ哲学にあった原理であるが，14 世紀の哲学者・神学者の William of Ockham が神学論争の上で多用したことで有名になり，オッカムの剃刀と呼ばれるようになった．つまり「同じことを説明しようとするなら，複雑な仮説よりも単純な仮説のほうがよい」がケチの原理であり，オッカムの剃刀である．

なお，データサイエンスにおけるモデルは，物理モデルや経済モデルのように確立された法則を表すだけのものとは限らない．データから何か新しいことを発見する助けとなるレンズの役割を果たすモデルから，法則というまでは成熟していないが，発見した関係や事実を汎用に記述する手段としてのモデルに至るまで，データサイエンスでは，さまざまなタイプのモデルが登場する．

図 6.1 は，現象，データ，人間，モデルの間の関係を表したものである．現象からデータ，そして人間に至る矢印で示されているように，人間は現象をデータを介して理解するしかない．しかし，データは現象のある側面を語っているにす

ぎない．人間は，そこから創り出されるモデルによって，現象の，ある側面を近

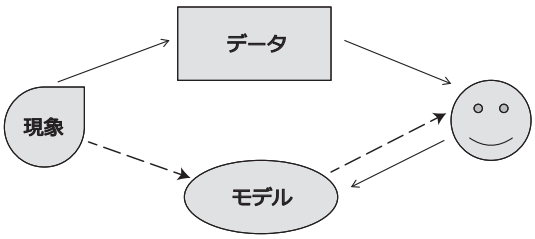

図 6.1 データからモデルを創り出し，それを介して現象を理解する．

似的に理解できるだけであることは，データサイエンスの一つの基本として，理解しておく必要がある．

データからモデルを創り出す手順，つまり図 6.1 の人間からモデルへの矢印で示された**モデリング** (modeling) は，次のような 4 ステップから成る．

1. 現象のどの側面をモデリングしようとするのか，目的を絞る．
2. そのために必要な情報がデータにどの程度含まれているのかを見極める．
3. 創り出したモデルが，どの程度の近似になれば十分なのか，その水準を設定する．
4. モデルを創り出したら，そのモデルがどれだけ役立つか，理解が容易かを，冷静に判断する．

2 番目のステップの「必要な情報がデータにどの程度含まれているか」は定性的な表現でしかないが，データ解析をしていると，どうしても**データの質** (data quality) が気になってくる．

データの質

データの質は，そのデータが現象のさまざまな側面を，どれだけよく反映しているかで決まる．したがって，ある側面では質の高いデータでも，別の側面では質が低いこともありうる．そのデータが，どんな側面からみても現象をうまく反映してないならば，それはゴミでしかない．「うまく反映」を厳密に定義するのは難しいが，大雑把にいえば，データが，現象をどれだけ偏りなく反映しているか，どれだけ正確に反映しているかである．たとえ，大量の記録があっても，偏りが大きければ，それを補正する確かな手段が見つからない限り，その価値は大幅に減ずる．

さまざまな角度からいくら解析しても何も見えてこない．そんなときは，適当な結果でこじつけるよりは勇気をもって「このデータからは何も確かなことは言えない」ということが大切である．[10] つまり「このデータから何々が言えた」と同じくらい「このデータからは何も確かなことは言えない」という結論も価値がある．ただし，**どうして何も言えないのか**について合理的な説明は欠かせない．それがなければ誰も納得しないからである．しかし，その説明さえはっきりしていれば，その後の解析の方向を再検討したり，より情報量の多いデータを取り直したり，探したりする契機となり，それまでの努力は無駄にならない．

10) カリフォルニア大学の John Rice 教授に「このデータからは何も言えないという結論も重要な結論だよね」と言われ，同感だったことを思い出す．

モデル

一般的に，モデルは本物をそのまま扱うのが困難なとき，代わりの役目を果たすものである．データサイエンスのモデルは，すでに確立した法則を表すものだけとは限らない．データから何か新しいことを発見する助けとなる，レンズの役割を果たすモデルから，法則というまでは成熟していないが，発見した関係や事実を汎用に記述する手段としてのモデルに至るまで，さまざまなモデルが登場する．

6.4 データサイエンス

データサイエンス (data science) は「データを対象とした科学」である．科学は「なぜ」という疑問から出発し，その答えを見出すことに努力を重ねる．また，その過程で，なんらかの発見があることも期待している．したがって，データサイエンスは，「抽象的なデータ」というものを対象とするなら，データをどう認識すればよいか，どう扱えばよいかといった，一般的な「なぜ」に答えることになるし，「具体的な個々のデータ」を対象とするなら，その背後にある現象に関する「なぜ」に，どのようにしたら，データを通した答えを見つけられるかも研究することになる．したがって，データサイエンスは

データから新たな価値を創出する科学

と定義するとよいだろう．[11] 新たな価値の中には，発見とまではいかないような，数々の「なぜ」という疑問に対する答えも含まれる．データから生まれた「なぜ」に対する答え，つまり原因を見出そうとするのは，なにも知的好奇心からだけではない．何かおかしなことが起きたとき，原因さえつかめていれば，そこに変化があったのではないかと調べることで，すぐ対処できるが，原因がつかめていなければ手の打ちようがない．後の祭りである．

11) これとほぼ同じデータサイエンスの定義が Wikipedia 英語版にも見受けられる．

データサイエンスに対比するものとして**データエンジニアリング** (data engineering)，つまり「データの工学」があるが，一般的に，工学は何か目標を定め，それを「実現する」ことが目的である．それに対し，基本的に，科学は「なぜ」という疑問に対する答えを見つけることが目的である．こうした，科学と工学の立場の違いは，あまり表立って議論されることは多くないが，データに取り組むときの姿勢には大きく影響する．

たとえば，データから，売上増につながる販売方法を見つけたいとしよう．データサイエンスの実践では，まず，データから，何を (What)，なぜ (Why)，どこで (Where)，いつ (When)，だれが (Who) といった 5W の疑問に対する答えをデータから見つけた上で，1H (How) あるいは 4H(How, How much, How many, How long) に対する答えを見つけようとする．[12)]

これに対し，データエンジニアリングでは，そのような疑問に一つひとつ答えていくのではなく，一直線に，「データ」という入力を与えると「すこしでも売り上げ増につながる販売方法」を出力してくれるような，何らかのシステムをつくりあげることを目標とする．たしかに，このアプローチは，とりあえず何らかの結果が得られるものを提供できる，という意味では便利で可用性も高いが，どこまで有効な販売方法なのか，他にはないのか，どのような状況で有効なのか，といった数々の「なぜ」に答えることは難しい．[13)] たとえば，**機械学習**した結果から，なぜそれがうまく働くのか，その理由を探しだそうと思っても明確な答えを見出すのは困難なことが多い．

12) 5W1H あるいは 5W4H は，新聞記事などを書くときの欠かせない 5 大要素として取り上げられることが多いが，データを解析するときにも，これらの要素を頭に置いてデータをブラウジングするとよい．

13) エンジニアリングでは，しばしば “It works.” が錦の御旗になる．モデリングの基本である**ケチの原理**とは対照的である．

機械学習

機械学習 (machine learning) は，コンピュータ上のアルゴリズムにデータを与え，そのデータに応じた適切な結果を出力するように，アルゴリズムを変化つまり学習させるエンジニアリングである．神経細胞ネットワークを模したアルゴリズムのニューラルネットワーク，遺伝子の継承を模したアルゴリズムの**遺伝的アルゴリズム** (genetic algorithm) などがある．最近では，ニューラルネットワークを 3 層以上積み重ねた**深層学習** (deep learning) も注目を集めている．なお，確率的ニューラルネットワークによる TOPIX の予測例 [22] も参考になるに違いない．

しかし，データサイエンスの実践とデータエンジニアリングに，そんなに大きな違いがあるわけではない．データサイエンスの実践 [14)] でも「なぜ」に答えることで，データの背後にある現象の全体像をつかみ，それをモデルの形で表現し，具体的な価値を創出しようとする．上の例でいえば，売上増につながる販売方法

14) あえて「データサイエンスの実践」としたのは，サイエンスはあくまでもサイエンスであって，それを実際に適用するのとは，少し意味合いが異なるからである．

を，**モデルによる説明とともに**具体的に示すのがデータサイエンス実践のゴールであり，売上増につながるような何らかの販売方法を出力するようなシステムを作り，それを**動かして見せる**のが，データエンジニアリングのゴールである．しかし，そのシステムを動かして見せる段階は，モデルがシステムに置き換わるだけで，データサイエンスの実践そのものである．[15)]

データエンジニアリングには，これ以外にもさまざまな技術分野があり，データベース管理から，データから地球規模のシミュレーションモデルを構築する**データ同化** (data assimilation) の技術 [15] に至るまで，大規模で複雑に絡みあったデータを処理する技術は，データサイエンスを実践する上でも欠かせない．データサイエンスの実践では，「なぜ」という疑問に答えることを基本としながらも，必要に応じてこれらの技術も大いに活用していく必要がある．

近代統計学 (modern statistics) の父と呼ばれる R.A. Fisher (1890-1962) は，まさしくデータサイエンスを実践した人であった．彼は，常にデータから出発し，その背後にある現象や法則を明らかにしようと，さまざまな面からの解析を重ねた．しかし，近代確率論にもとづく推測理論が，数学としての成熟度を増すにつれ，数学の枠組みでさまざまな推測法を開発することに重きが置かれるようになった．その結果，統計的推測法の理論的裏付けを行う**数理統計学** (mathematical statistics) の興隆とともに，データの科学という出発点は忘れられがちになっていったのは自然の流れであろう．[16)]

そこに警鐘を鳴らしたのが，第 4 章で紹介した J. Tukey(1915-2000) である [8,50,51]．彼の提唱した**探索的データ解析** (EDA, exploratory data analysis) は，今になってようやく，データサイエンスの興隆という形で開花した．実際，2001 年になって**データサイエンス**を提唱した W.S. Cleveland [11] も J. Tukey の弟子の一人である．[17)]

図 6.2 は，現在のデータサイエンスに至るまでの時間的な流れを簡単に図示したものである．この図からわかるように，近代統計学の次に位置する数理統計学は，探索的データ解析を生み出す一方，品質管理や薬効検定など社会的な制度も生み出した．一方，数理統計学とは別の流れとして，**記述統計** (descriptive statistics) がある．記述統計は，データを集計し，それを他人にわかりやすく説明するための説明技法である．本書の第 I 部で紹介したさまざまな代表値や分布状態の表現法なども，この記述統計に属する．

ビジネス分野では，記述統計が，データを企業戦略に役立てる **BI** となり，さらに「データの中に有意なパターンをみつけ，それをうまく他人に伝える」**アナリティックス**に進化した．[18)]

ここまでくると，もうデータサイエンスにかなり近い．またその目的は，探索的データ解析とも近い．実際，データサイエンスの実践も，常にこのような

15) 日本では，統計学を「方法の学問である」とし，頭のなかで，目新しい方法を考えたり，外国から輸入することこそが，仕事と思っている学者も多いが，データを忘れた統計学は，もはや抜け殻である．

16) 事前情報を事前確率という形で導入する**ベイズ統計学** (Bayesian statistics) も，その客観性に対する疑問から数理統計学の主流にはなり得なかったが，最近，サーチエンジンへの応用で再び見直されている．

17) W.S. Cleveland はデータサイエンスのアイディアを，1996 年に国際分類学会出席のため来日したとき，著者から得たようである．このとき著者は同時開催の日本統計学会年会で共通テーマとして「データサイエンス I, II, III」を企画していた．

18) 本章の冒頭で引き合いに出した**解析学**の英語は analytics でもあるが，ここでのアナリティックスと解析学の間に直接の関係はない．

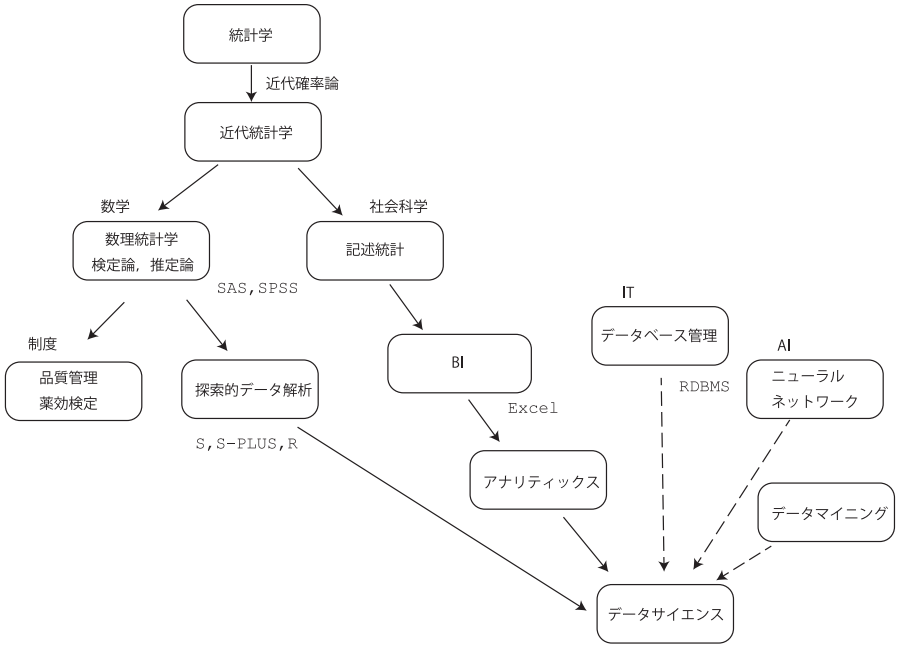

図 6.2 データサイエンスへの流れ

「データの中に有意なパターンを見つける」ことから始まる．違いは，そのあと「単純な要素に分解し，それを再び統合する」かどうかであり，見つけたことを「モデル」という形で抽象化して他人に伝えるかどうかである．

このような歴史的な流れ以外に，データサイエンス成立には，データに関連したさまざまな分野の発展も大きくかかわっている．コンピュータサイエンス一般の発達はもちろんのこと，その中でも IT (information technology) [19] と **AI** (artificial intelligence) が密接な関係にある． いずれも，データエンジニアリングの一部と考えられるが，IT 技術の中でも， 第 I 部でも紹介した関係形式データベース管理システム (RDBMS) の基礎をなす関係形式は，基本的なデータのとらえ方として重要であり，データサイエンスの基本でもある．これに関しては，6.7 節でさらに詳しく説明する．

一方，人工知能という夢を追いかける AI の分野では，自動翻訳や音声認識を具体的な目標として**ニューラルネットワーク** (neural network) による学習や**ベイジアンネットワーク** (Bayesian network) [20] による音声認識などで数々の成功を収めてきた． これらの技術は，一般的にデータから何かを見出すのにも使えると期待されている．また，**データマイニング** (data mining) は「大量のデータ

19) IT は最近 **ICT** (information and communication technology) に名前を変えた．情報を処理するのに通信技術が欠かせない要素となってきたからである．

20) ベイジアンネットワークの簡単な例が第 8 章にある．

統計学とデータサイエンス

図 6.8 で，Statistics が Data Science の一部となっていることに疑問を持たれる方もいらっしゃるかもしれない．すでに少しでも統計学を学び，本章を読み進めてきた方ならば，その違いはある程度おわかりになるであろう．データサイエンスの出発点は「データをどうとらえるか」であり，ゴールは「データから宝を見つけ出すこと」である．一方，大学などで教わる統計学は，図 6.2 でいうところの数理統計学や記述統計学である．

数理統計学は確率論をベースにし，確率変数とその観測値という定式化ですべての話が進む．しかし，データサイエンスがカバーする範囲で，このような定式化が有効なのは，図 6.3 の「解析，モデル構築」に到達したときの，しかもごく限られた場合である．それどころか，形式的な有意性や p **値** などにとらわれすぎて，せっかくの宝を見失う恐れのほうが大きい．また，記述統計学は，第 I 部で紹介したようなデータの代表値を求めたり図示したりする技法に留まる．

本章の後半は，統計学を超えたデータサイエンス固有な部分であり，次章以降は，統計学と重複する部分もあるものの，データサイエンスの実践という視点で見直している．統計学の看板をデータサイエンスに架け替えるだけでは，宝はみつからない．読者には，ぜひデータサイエンスを的確に理解し，未来を切り開いていただけるものと期待している.[49)]

[49)] 著者がデータサイエンスを提唱し始めたとき，データサイエンスは統計学の一部と思っていた学会員も多かった．いまでも，そう思っている，あるいはイコールと思っている方も多いのではなかろうか？

データサイエンティストへの道

データサイエンティストはデータのプロとして，さまざまな分野の人々と協力して，仕事を進めることになる．そうすると，データに関する知識や経験，モデリングのスキル，コンピュータを駆使する能力などはもちろんのこと，適用分野に関する知識や経験も身に着ける必要がある．さらに，広い範囲のコミュニケーション能力，外国語のリテラシー，ビジネス感覚や経済状況を見通す力，経営のセンスなど，きわめて多岐にわたる素養が問われることが多い．したがって，それぞれの分野の専門家の 3〜4 倍の勉強が必要になることは覚悟しよう.[50)]

[50)] 編集者から，**心身の健康**もデータサイエンティストの要件の一つとして加えるべきだという意見が寄せられた．どんなことをするにしても，まずは心身の健康が大切なことは確かである．

6.6 データファイル

データを解析するとき，コンピュータを使わないことは，まずない．しかし，はじめから，データが関係形式のような，きれいなファイル形式になっていると

6.9.3 汎用なデータ解析環境 S と R

1980 年代に米国でデータ解析を柔軟に行える環境を一から作り直そうという気運が盛り上がり，ベル研究所の J.M. Chambers と R.A. Becker の二人によって「データ解析とグラフィックスのための対話型環境」が創られた．これが最初の **S** である [5]．[89] その頃ベル研究所では，新しいオペレーティングシステム UNIX やプログラミング言語 C が同じように一から作り直されていた．S にはその開拓者魂が引き継がれている．言語としては未熟であった，この S にオブジェクト指向を導入し，本格的な言語環境として 1988 年にリリースされたのが現在の R や S-PLUS のもとになる S である [5,6,9,43]．その過程で著者もその改良に協力している．

ベル研究所は米国の電信電話会社 AT&T 傘下の研究所であったため，その研究成果を広く公開する義務があった．この S も広く公開され，一つは商用版の **S-PLUS** (`http://www.msi.co.jp/splus/`)，もう一つはパブリックドメイン版の **R** (`https://www.r-project.org/`) として発展してきた．しかし，オン・ディスクとオン・メモリーという違いはあるものの，どちらも基本的に S を踏襲しており基本的な違いはない．もちろん，S-PLUS は商用版として信頼性に重きを置き，**ビッグデータ**を扱いやすくする拡張なども早い時期から手掛けており，R は世界中のユーザがボランティアとして自由で活発な改良を重ねてきた．[90] ユーザ作成の膨大な公開ライブラリを有するのも大きな特徴である．そのため，たとえばゲノム解析の分野では R 上でこの公開ライブラリを使うのが一つの標準となっている．

このソフトウェアの最大の特徴は，その柔軟性と汎用性にある．最初に説明したようにデータから新たな価値を創出する過程は一本道ではなく，様子を探りながら試行錯誤を繰り返す，探索の旅である．この旅を成功裏に終わらせるためには，必要な道具がいつでも気楽に快適に使えるような柔軟な環境が必要となる．S ではこれを関数型言語とオブジェクト指向の組合せで実現している．関数型言語ならば，基本的な道具を用意しておきさえすれば，それを組み合わせることでいくらでも必要を満たすことができ，結果の再利用も容易である．また，オブジェクト指向によって，解析対象の多様性をあまり意識せずに一般的な操作ですませられる．S は探索的データ解析を提唱した J. Tukey の弟子たちが作っただけあって，30 年以上たった今のデータサイエンスの時代にも色あせずに広く用いられるソフトウェアとなっている．

しかし，本家の S は 1990 年に入って開発が終了しており，開発者自身が考えていた数々の改良点も手つかずのまま残っている．一つはグラフィックス機能の貧弱さである．これは，データ解析のためには，コンピュータの負荷が軽い必要

89) 彼らは「S は Statistics の頭文字ではない．Statistics はソフトウェアにとってはいわば kiss of death で，統計解析用のソフトウェアというだけで，普及を妨げてしまう恐れもある」といって，それまでの統計解析ソフトウェアとは一線を画し，一から開発した．S が何の頭文字なのか，その答えは [43] の p.46 にある．ちなみに，1995 年にリメークされた映画『死の接吻 (Kiss of Death)』を念頭に置いた発言である．もともと，この英語は「命取り」あるいは「災いの元」という意味で，いちど統計解析用ソフトウェアとレッテルを張られたらもうお終い，という危機感が，こう言わせた．

90) Rtools というフリーソフトウェアを使うと R の使用環境はいくぶん改善する．

8.1.1 放射性物質拡散データ

2011 年 3 月 11 日に起きた東北地方太平洋沖地震が引き起こした津波は，家屋や施設を押し流しただけでなく，福島原子力発電所の事故を引き起こし，大量の放射性物質が全国に拡散するといった大きな災害をもたらした．本節では，どのように放射性物質が拡散していったのか，公開データからわかる範囲で見てみよう．[10)]

原子力規制委員会放射線モニタリング情報の Web サイト

```
http://radioactivity.nsr.go.jp/ja/
```

からダウンロード [11)] した，47 都道府県の市区町村の施設等で 3 月 11 日 15 時から 19 日 23 時まで 1 時間ごとに測定された**空間放射線量率** (air radiation dose rate) データの CSV ファイルが，`radiation.csv` である．これを R に読み込み，その一部を眺めてみると

```
> Radiation=read.csv("radiation.csv")
> Radiation[1:3, 1:10]
   place mplace X1h X2h X3h X4h X5h X6h X7h X8h
1 北海道 札幌市  NA  NA  NA  NA  NA  NA  NA  NA
2 青森県 青森市  NA  NA  NA  NA  NA  NA  NA  NA
3 岩手県 盛岡市  NA  NA  NA  NA  NA  NA  NA  NA
```

のようになる．[12)] ファイルには，観測施設ごとの記録が 1 行ずつ入っており，第 1 列は都道府県名，第 2 列は観測施設の所在地，第 3 列以降が 3 月 11 日 15 時から 1 時間おきの空間放射線量率 (μSv/h，マイクロシーベルト毎時）になっている．3 月 13 日になるまでは，放射線は観測されていないので値はすべて `NA` である．

このデータフレーム `Radiation` は，エンティティーを都道府県（観測施設）に想定したデータテーブルになっている．このデータが各施設ごとの観測データを集めたものである以上，このようなエンティティーを想定するのは当然である．しかし，いざデータを取得してみると，空間放射線量率がどのように時間変化していったかに興味の中心が移る．つまり，時刻をエンティティーにして，各時刻での 47 都道府県での空間放射線量率という見方をしたくなる．[13)]

そこで，このデータテーブルを転置し，データ行列に直す．その際，第 1 列の県名はデータ行列の列ラベルに移す．第 2 列の所在地は，とりあえずこれからの解析には必要ないので，除いておくことにする．

```
> Radiation2=t(as.matrix(Radiation[, -(1:2)]))
> colnames(Radiation2)=Radiation[,1]
> Radiation2[1:3,1:8]
      北海道 青森県 岩手県  宮城県 秋田県  山形県 福島県 茨城県
```

10) 慶應義塾大学での 2011 年秋の学部 3 年生向けのゼミの中で，菅澤翔之助君らとともにおこなった解析をもとにしている．

11) 2014 年末までは，この Web サイトからダウンロードできたが，現在どこからダウンロードできるか不明である．

`radiation.csv`{DSC}

12) 読み込んだとき，`X1h` のような，頭に `X` が付け加わったラベルになるのは，R では数字で始まる名前は許されないからである．

MacOS 上の R では，データをファイルから読み込むとき，ファイルの漢字コードと R での漢字コードを合わせるため `fileEncoding="cp932"` のような追加引数が必要になるかもしれない．

13) このように，同じデータであっても，何をエンティティーとして想定するかは，目的によって変わってくる．

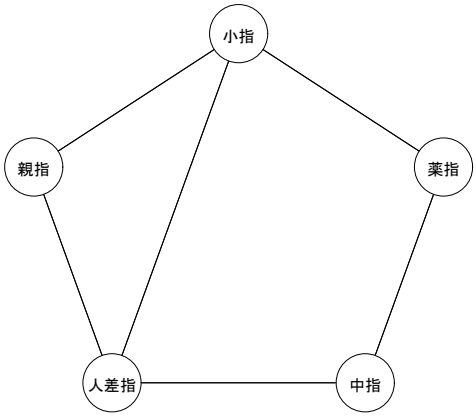

図 9.4 5 指の長さのグラフィカル表現

ぶかどうかを条件付き独立性で定めるのではなく，各頂点（変量）について，祖先の値を条件とする条件付き確率が親の値だけで定まる，いわゆる**マルコフ性** (Markov property) を持つ有向グラフになるように結ばれる．特に非循環型の有向グラフになっている場合**ベイジアンネットワーク** (Bayesian network) [15] と呼ばれ，特に離散値をとる変量の場合によく利用されるグラフィカル表現である．

[15] ベイジアンネットワークと呼ばれる理由は，このネットワークでの様々な条件付き確率を求めるのに，**ベイズの公式** (Bayes formula) が役立つからである．ちなみに，ベイズの公式をベイズの定理と呼ぶ人もいるが，これは定理というほどのものではなく簡単な公式にすぎないので，名前に惑わされないようにしてほしい．

ベイズの公式

ベイズの公式は，条件付き確率の「条件と結果の役割を交換」する公式で，高校での事象の確率の形で書けば，事象 A と B に関する条件付き確率の

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

のような書き換えにすぎない．さらに $A_i, i = 1, 2, \ldots, m,$ が排反で和が全事象となるなら，

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{\sum_{j=1}^{m} P(B|A_j)P(A_j)}$$

のように，条件付き確率は事象 B を条件から結果に移した条件付き確率と，事象 $A_i, i = 1, 2, \ldots, m,$ に関する確率だけで書き表せる．

第10章　確率モデル

本章では，確率というレンズを通してデータを眺め，現象をモデル化するのに，**確率モデル** (stochastic model) [1] がどのように役立つか，地震データと株価データを例として，その役割を説明する．この2種類のデータは，まったく異なる性格のデータでありながら，同じようなアプローチで，その裏に潜む現象を記述できる．これが確率モデルの持つ汎用性である．さらに，第6章で紹介した，真鯛放流捕獲データの確率モデルのさらなる高度化，呼吸量心拍データの心拍間隔の分布の検討なども行う．

[1] ここでいう確率モデルは，確率が本質的な役割を果たすような現象に対するモデルである．第8章の回帰モデルの誤差項に対する，正規分布のような便宜的なモデルとは意味も役割も異なる．

10.1　地震データ

日本は地震国であり，いつ大きな地震に襲われるかわからない．いつ起きるか

図 10.1　2011 年 12 月 3 日気仙沼にて（著者撮影）

予測できたらどんなに助かるか，国民すべてに共通な願いであろう．残念ながら天気予報に匹敵するような精度での予測はまだできないが，それでも気象庁や大学，研究所を中心として精力的な観測と研究が続けられている．ここでは，デー

10.5.4 地震の予測

地震の発生をポアソン現象としてモデリングし，予測することには限界がある．ポアソン現象は，背後に特別な発生メカニズムを想定していないからである．現在，主に予測に用いられているモデルは **BPT モデル** (Brownian passage time model) [28] である．これは，プレートに蓄積したひずみがある限界に達するとそのエネルギーが地震として放出されるという考えにもとづいている．

時刻 t におけるひずみの蓄積 $X(t)$ を

$$X(t) = at + \sigma B(t), \quad t \geq 0, \quad B(0) = 0$$

でモデル化すれば, $X(t)$ が初めてあるレベル x を超えた時点 $T = \inf\{t \,|\, X(t) \geq x\}$ が地震の発生時点になる．このモデルでは，ひずみは基本的には時間の経過とともに a の率で蓄積するが，それだけでは説明しきれない部分がブラウン運動でモデリングされている．

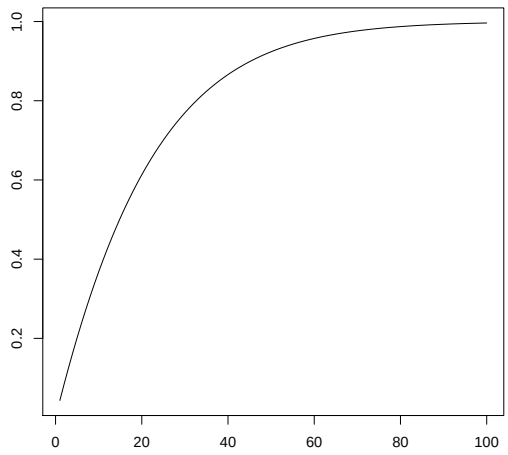

図 10.13 東海地震の発生確率

歴史的には，T の分布は古くから**逆ガウス分布** (inverse Gaussian distribution) あるいは**逆正規分布** (inverse normal distribution) として知られていた．[33] この名前は，増分が正規分布つまりガウス分布に従って動いているブラウン運動の値が，レベル x に最初に達する時点の分布であることに由来しており，その確率密度は

$$f(t) = \frac{|x|}{\sqrt{2\pi\sigma^2 t^3}} \exp\left\{-\frac{(x-at)^2}{2\sigma^2 t}\right\}, \quad t \geq 0 \tag{10.15}$$

[33] 地震学分野では, 逆ガウス分布は BPT 分布と呼ばれている．

で与えられる．

このモデルは単純ではあるが，パラメータ a, σ, x だけで定まるので，限られた過去の地震の記録から推測せざるをえないような状況では，その可用性は高い．このモデルは，一度地震が起きればひずみは 0 に戻る，と考えているので，時間の原点は地震が最後に起きた時点に設定される．したがって，現時点を t_0 とすれば，s 時間後までに地震が起きる確率は

$$\mathrm{P}(T \le t_0 + s | T > t_0) = \frac{\int_{t_0}^{t_0+s} f(u)du}{\int_{t_0}^{\infty} f(u)du} \tag{10.16}$$

として求まる．

あと 30 年以内に東海地震が起きる確率は 87%?

あと 30 年以内に東海地震が起きる確率が 87%と 2002 年に発表されている [33]．これから 1 年当り 2.9%，1 か月当り 0.2% と考えるのは早計である．もちろん，各年で一定の確率で独立に起きるとして，$1-(1-p)^{30} = 0.87$ を解いて，1 年に起きる確率は $p = 0.066$ つまり 6.6% とするのも正しくない．きちんとモデルにもとづいて求める必要がある．確率密度関数 (10.14) は，期待値 x/a を μ，形状パラメータ $\lambda = x^2/\sigma^2$ と置けば， 逆ガウス分布 $\mathsf{IG}(\mu, \lambda)$ の標準形

$$f(t) = \sqrt{\frac{\lambda}{2\pi t^3}} \exp\left\{-\frac{\lambda(t-\mu)^2}{2\mu^2 t}\right\},\ t \ge 0$$

に書き換えられる．報告 [33] では，過去の記録から $\alpha = \sqrt{\mu/\lambda} = 0.24$，平均活動間隔 μ を 118.8（年）として 2002 年 1 月 1 日時点での次に大地震の起きる確率を求めている．前回の東海地震は 1854 年 12 月 23 日だったので $t_0 = 147.02$（年）である．これらの条件で求めた地震の発生確率 (10.15) を，横軸を年，縦軸をその時までの発生確率に取ってプロットしたのが図 10.13 である．

```
> library(statmod)
> Bpt()
```

ただ，この計算だと 2002 年 1 月 1 日から 30 年以内に地震が起きる確率は 76.96%であり，報告書 [33] に掲載されている図から割り出した 87%と一致しない．実際に計算に用いた α が，報告書にある 0.24 ではなくもう少し低い値なのかもしれない．

pinvgauss(){statmod},
Bpt{DSC}

索 引