

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。

本書は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まい。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出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、影付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ある中に置かれている。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方丸佑紀君、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 11 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、ようやく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1章 データ	3
1.1 変数と変量とデータ	5
1.2 関係形式	8
1.2.1 湿度吸収実験データ	10
1.2.2 複式簿記データ	13
1.3 データの代表値	16
1.3.1 最小値, 最大値	18
1.3.2 平均値	18
1.3.3 標準偏差	21
1.4 偏差値	24
第2章 データ分布	27
2.1 1次元散布図	27
2.2 度数分布表とヒストグラム	29
2.3 度数分布多角形	32
2.4 ひとつと	33
第3章 データ分布の代表値	35
3.1 データ分布の中心	35
3.1.1 平均値	35
3.1.2 中央値	37
3.2 データ分布の広がり	41
3.2.1 標準偏差	41
3.2.2 平均絶対偏差	42
3.2.3 四分位数	43
3.3 データ分布の要約値	44
3.4 データ例	45
3.4.1 春の訪れ	45

3.4.2	音楽アルバム	48
第4章	箱ひげ図	51
4.1	ネットワークの応答速度	51
4.2	箱ひげ図とロウソク足	54
4.3	ひとつこと	59
第5章	2変量データ	61
5.1	変量の相関関係	61
5.2	2元度数分布表と2次元散布図	63
5.3	2変量同時分布の代表値	64
5.3.1	共分散と相関係数	64
 第II部 データサイエンス		
第6章	データサイエンス入門	69
6.1	データに語らせる	70
6.2	ストラテジー	70
6.3	モデル	72
6.4	データサイエンス	75
6.4.1	データの上流から下流まで	79
6.4.2	データリテラシー	89
6.5	データサイエンティスト	95
6.6	データファイル	97
6.6.1	フラットファイル	98
6.6.2	マークアップファイル	99
6.6.3	バイナリーファイル	100
6.7	関係形式データベース	101
6.7.1	テーブルに対する演算	101
6.7.2	SQL	102
6.7.3	キー	103
6.7.4	ドメイン	107
6.8	関係形式データベースを超えて	108
6.8.1	2次データ	108
6.8.2	テーブル間の関係	112
6.8.3	さまざまなレベルでの属性	114
6.8.4	DandD	115

6.9	ソフトウェア	119
6.9.1	表計算ソフトウェア	120
6.9.2	統計解析ソフトウェア	120
6.9.3	汎用なデータ解析環境 S と R	121
6.10	データ行列と線形代数	122
6.10.1	データ行列	123
6.10.2	個体空間と変量空間	123
6.10.3	中心化	124
6.10.4	尺度規準化	125
第 7 章	個体の雲の探索	129
7.1	クラスタリング	129
7.1.1	ウイルス RNA 変異データ	132
7.2	主成分分析	142
7.2.1	新しい座標軸の求め方	143
7.2.2	都道府県之力	148
7.2.3	特異値分解	153
7.3	高次元個体空間の可視化 —TextilePlot—	156
第 8 章	変量間の関係	167
8.1	回帰モデル	167
8.1.1	放射性物質拡散データ	173
8.1.2	湿度吸収実験データ	179
8.1.3	脊柱後弯症データ	183
8.1.4	真鯛放流捕獲データ	186
8.2	非線形回帰モデル	187
8.3	正準相関分析とコレスポンデンス分析	188
8.3.1	正準相関分析	188
8.3.2	コレスポンデンス分析	193
第 9 章	変量間の相関	199
9.1	相関係数と偏相関係数	199
9.2	相関と独立性	205
9.3	偏相関と条件付き独立性	210

第 10 章 確率モデル	215
10.1 地震データ	215
10.1.1 マグニチュードの分布	216
10.1.2 地震の発生間隔の分布	229
10.2 真鯛放流捕獲データ	232
10.3 株価収益率データ	236
10.4 心拍データ	239
10.5 ひとつこと、ふたこと	241
10.5.1 連続収益率と複利	241
10.5.2 正規分布	242
10.5.3 ブラウン運動	243
10.5.4 地震の予測	244
10.5.5 多変量正規分布	246
10.5.6 ガンマ分布	247
参考文献	249
索引	252

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。

main : 2015/11/12(15:42)

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という。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。

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、次の解析の戦略もはっきりさせられる。第8章は変量間の関係である。変量間の関係をつかむことが、よりの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も、数学は苦手と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、それを避けることで、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方が大きい。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、強し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⁾

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を探り、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。

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。

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。

索引

あ

R 121
 RSA (respiratory sinus arrhythmia) 80
 RNA シークエンス (RNA sequence) 133
 RDBMS(relational database management system) 8, 102, 115
 ICD(international classification of diseases) 109
 ICT(information and communication technology) 78
 ID (identification) 4, 84, **103**
 ID 型 (id) 12
 赤池情報量規準 (Akaike's information criterion) 189
 値空間 (range space, image space) 155
 値の制約 (value restriction) 114
 値の表記 (value representation) 113
 アナリティックス (analytics) 77, **120**
 アライメント (alignment) 134
 安定分布 (stable distribution) **237**, 238

い

ER モデル (entity-relationship model) ... 104
 EDA(exploratory data analysis) 57, **77**
 位置 (location) 23
 — 移動 (shift) 20
 — 不変 (invariant) **23**, 146
 1 次元散布図 (one dimensional scatter plot) 28
 1 次データ (primary data) 48, **80**
 位置尺度共変性 (location-scale equivariant) 20

一般化加法モデル (generalized additive model) 189
 一般化線形モデル (generalized linear model) 163, **183**
 遺伝的アルゴリズム (genetic algorithm) .. 76
 移動平均 (moving average) 187
 意味 (semantic) 107
 因子データ (factor data) 12
 インスタンス (instance) 104
 インターデータベース (inter database) .. 116

う

Web ページ (web page) 100

え

AI (artificial intelligence) 78
 HTML (hyper text markup language) 99
 エージェント (agent) 116
 S 121
 SQL 102
 SPSS 120
 S-PLUS 121
 XML(extensible markup language) 100
 NA (not available) 85
 NCBI(National Center for Biotechnology Information) 133
 演繹 (reduction) 7
 円グラフ (pie chart) 89
 エンティティ (entity) **103**, 104

お

オッカムの剃刀 (Ockham's razor) 73
 オッズ (odds) 184
 オブジェクト (object) 104
 オブジェクト指向言語 (object oriented language) 85

か

回帰 (regression)
 頑健 (robust) — 172
 —関数 (function) 167
 —直線 (line) 169
 —モデル (model) 167
 回帰樹 (regression tree) 189
 階級 (class, bin) 29
 —値 (value) 30
 —幅 (size) 30
 階数 (rank) 155
 解析学 (analysis) 69, 77
 階層的クラスタリング (hierarchical clustering) 131
 外部キー (foreign key) 112
 核空間 (kernel space) 155
 拡張ドメイン (extended domain) 113
 核補空間 (kernel complementary space) 154, 155
 確率的フラクタル (stochastic fractal) 219
 確率点 (quantile) 206, 222
 確率分布関数 (probability distribution function) 206
 確率変数 (random variable) 206
 —の内積 (inner product) 208
 —のノルム (norm) 208
 確率密度関数 (probability density function) 206
 確率モデル (stochastic model) 215

貸方 (credit) 13
 加速度 (acceleration) 27
 カテゴリ型 (category) 12
 カテゴリカルデータ (categorical data) 12, 193
 カテゴリ変量 (category variate) 12
 過分散 (over dispersion) 91
 可変長フィールド形式 (variable length field format) 99
 カラム (column) 9, 81, 98, 101, 102
 借方 (debit) 13
 刈込平均値 (trimmed mean) 39
 刈込率 (trimming proportion) 39
 関係 (relation) 9, 167
 関係グラフ (relation graph) 9
 関係形式データベース (relational database) 8, 101

頑健回帰 (robust regression) 172
 慣性モーメント (moment of inertia) 41
 ガンマ分布 (gamma distribution) 247

き

キー (key) 103
 自明な (trivial) — 103
 —の既約性 (irreducibility) 103
 機械学習 (machine learning) 76
 刻み (tick marks) 87
 擬似距離 (pseudo distance) 131
 記述統計 (descriptive statistics) 77
 基数系 (radix system) 114
 期待値 (expectation) 207
 輝度 (brightness) 45, 47
 帰納 (induction) 7
 逆ガウス分布 (inverse Gaussian distribution) 244
 逆正規分布 (inverse normal distribution) 244
 CAPM (capital asset pricing model) 238
 Q-Q プロット (quantile-quantile plot) ... 223

254 索引

強度 (intensity) 230
 共分散 (covariance) 64
 共変性 (equivariant) 40
 行列のノルム (matrix norm) **127**, 146
 局所回帰 (local regression) 187
 距離 (distance)
 一致 (coincidence)— 130
 街区 (cityblock, Manhattan)— 130
 最大 (maximum)— 130
 集合間の— 131
 絶対和 (absolute sum)— 130
 ユークリッド (Euclidean)— 130
 寄与率 (contribution) 145
 記録 (record) **8**, 98
 記録数 (the number of records) 18
 近代統計学 (modern statistics) 77

く
 空間放射線量率 (air radiation dose rate) 173
 Gutenberg-Richter 則 (Gutenberg-Richter law)
 216
 区切り記号 (separator) 99
 くびれ (notch) 53
 組換え (recombination) 132
 クラスタ (cluster) **83**, **130**
 クラスタ樹 (cluster tree) **135**, 160
 クラスタリング (clustering) 129
 グラフィカル表現 (graphical representation)
 213
 クリーニング (cleaning) 72
 クリッピング (clipping) 48
 クレンジング (cleansing) 72

け
 計画行列 (design matrix) 168
 経験分布関数 (empirical distribution function)
 221

k -means 法 (k -means clustering) 131
 ケチの原理 (principle of parsimony) **73**, 234
 欠損値 (missing value) 18, 85, 113, **114**
 現象 (phenomenon) 69
 —のモデル化 83
 ケンドールの順位相関係数 (Kendall's rank correlation coefficient) 205

こ

公差 (tolerance) 18
 降順 (descending order) 37
 後進差分 (backward difference) 59
 誤差項 (error) 167
 五数要約 (five number summary) 44
 個体 (individual) 123
 —の雲 (cloud) 129
 個体空間 (individual space) ... **123**, 124, 144
 固定欄形式 (fixed field format) 98
 固有属性 (native attribute) 106
 固有値 (eigenvalue) 145
 固有ベクトル (eigenvector) 145
 コレスポンデンス分析 (correspondence analysis) 193
 コントラスト (contrast) 47

さ

座位 (locus) 140
 最小二乗法 (least squares) 169
 最尤推定 (maximum likelihood estimate) 235
 SAS 120
 座標軸 (coordinate axes) 142
 —のクラスタリング (clustering) ... 160
 差分 (difference) **59**, 229
 サポートベクターマシン (support vector machine) 160
 残差ベクトル (residual vector) 169
 3 次データ (thirdly data) 112

算術平均 (arithmetic mean) 18
 散布図 (scatter plot) 17, **88**
 サンプルング (sampling) 71, **111**

し

CSV (comma separated values) 99
 視覚表現 (visualization) 28
 時間 (time) 85
 時系列 (time series) 83
 時系列図 (time series plot) 89
 自己資本収益率 (return on equity) 205
 指示関数 (indicator function) 131
 指数分布 (exponential distribution) 225
 実験計画 (design of experiments) 111
 実現幅 (realizable band) **91**, **224**
 実体 (instance) 103
 時点系列 (time points) 82
 四分位数 (quartile) 43
 四分位範囲幅 (interquartile range) 43
 四分位偏差 (interquartile deviation) 43
 射影 (projection) 102
 射影行列 (projection matrix) 125
 尺度 (scale) 6, **23**, 87
 — 共変 (equivariant) 23
 — 不変 (invariant) 147
 — 変換 (transformation) 20
 尺度規準化 (scaling) **126**, 147
 収益率 (return rate) 236
 重心 (barycenter) 35
 自由欄形式 (free field format) 99
 主キー (primary key) 103
 縮約値 (aggregated value) **44**, 109
 主成分 (principal component) 144
 — 行列 (matrix) 144
 — 軸 (axes) 144
 — 分析 (analysis) 142
 順位 (rank) 38
 瞬間死亡率 (instantaneous mortality rate) 233
 順序 (order) 38
 順序カテゴリ (ordered category) 162
 順序集合 (ordered set) 9
 順序統計量 (order statistics) 38
 条件付き (conditional)
 — 確率分布関数 (probability distribution function) 213
 — 期待値 (expectation) 212, **213**
 — 相関 (correlation) 212
 — 相関係数 (correlation coefficient) 212
 — 独立 (independent) 211
 — 箱ひげ図 (conditional boxplot) ... 209
 昇順 (ascending order) 37
 小数法則 (law of small numbers) 230
 情報公開法 (freedom of information act) 119
 処理対比 (treatment contrast) 180
 仕訳表 (journal) 14
 深層学習 (deep learning) 76
 信頼区間 (confidence interval) 224
 信頼性工学 (reliability engineering) 234

す

図 (figure) 16
 水準 (level) **12**, 193
 — の集まり (levels) 12
 垂線プロット (vertical line plot) **18**, **88**
 推定値 (estimate) 199
 水平性規準 (horizontality criterion)
 157, 158, 159
 数学的帰納法 (mathematical induction) 7
 数理統計学 (mathematical statistics) 77
 裾 (tail) 228
 — が厚い (fat) 228
 — が重い (heavy) **50**, **228**
 — が長い (long) 228

256 索引

スティルチェス積分 (Stieltjes integration) **207, 208**
 ストラテジー (strategy) 70
 スピアマンの順位相関係数 (Spearman's rank correlation coefficient) 205

せ

生起時刻 (occurrence time) 232
 正規分布 (normal distribution) 242
 斉次性 (homogeneity) 229
 正射影 (orthogonal projection) 125
 正準相関分析 (canonical correlation analysis) 188
 生存確率 (survival probability) 233
 脊柱後弯症 (kyphosis) 161
 セグメント (segment) 83
 切断 (truncation) 226
 説明変数行列 (explanatory variable matrix) 168
 説明変量 (explanatory variate) **104**, 167
 0 次データ (source data) 79
 線形回帰モデル (linear regression model) 168
 線形性 (linearity) 19
 線形変換 (linear transformation) 20
 前進差分 (forward difference) 59
 線プロット (line plot) 88

そ

層 (stratum) 111
 相加平均 (arithmetic mean) 20
 層化無作為抽出 (stratified random sampling) 111
 相関 (correlation) **61**, **199**
 — 係数 (coefficient) **65**, **199**
 — 係数行列 (coefficient matrix) 209
 — 表 (table) 63
 相乗平均 (geometric mean) 20

相対度数 (relative frequency) 33
 属性 (attribute) 5, 7, **9**, 101, 102

た

タイ (tie) 37
 対数オッズ (logarithmic odds) 183
 対数収益率 (log return rate) 236
 大数の法則 (law of large numbers) 222
 対数尤度 (log likelihood) 234
 ダイナミックレンジ (dynamic range) 44
 対比 (contrast) 158, **181**, 193
 — 行列 (matrix) 181
 代表値 (summary) 17
 タグ (tag) 99
 多項ロジットモデル (multinomial logit model) 185
 多次元尺度構成法 (MDS) 131
 多変量正規分布 (multivariate normal distribution) 246
 探索的データ解析 (EDA) 57, **77**
 短縮名 (short name) **11**, 84

ち

中央値 (median) **37**, **223**
 度数分布表から求めた — 40
 中心化 (centering) **124**, 146
 中心極限定理 (central limit theorem) 242
 中心差分 (central difference) 60
 直交行列 (orthogonal matrix) 145

つ

対散布図 (pairwise scatter plot) 157, **199**
 通日 (cumulative days) 85, **218**
 通秒 (cumulative seconds) 86

て

DandD(data and description) 100, 114, **115**, **116**

- インスタンス (instance) 115
- TSV(tab separated values) 99
- DK (don't know) 85
- DNA シークエンス (DNA sequence) 132
- データ
 - アイリス— **105**, 158
 - ウイルス RNA 変異— 132
 - 患者調査— 108
 - 呼吸量心拍— **80**, 117, 239
 - 地震— 215
 - 湿度吸収実験— **10**, 179
 - 脊柱後弯症— **161**, 183
 - 株価収益率— **56**, **236**
 - ネットワーク応答速度— 51
 - ピクセル輝度— 45
 - 複式簿記— 13
 - 放射性物質拡散— 173
 - ボール投げの実験— **6**, 16, 33
 - 真鯛放流捕獲— **89**, 186, 232
 - 民力— **148**, 164
 - 目と髪の色の一— 195
 - 指の長さ— 61
 - ランキング— 48
- データ (data) **5**, 9, 102
 - きたない (disorganized)— 72
 - きれいな (organized)— 72
 - の拡張 (extension) 115
 - の規準化 (normalization) 25
 - の雲 (cloud) **123**, **129**
 - の質 (quality) 74
 - の縮約 (aggregation) 115
 - の商品化 (commercialization) 119
 - の代表値 (data summary) 16
 - の広がり (dispersion) **21**, 145
- データエンジニアリング (data engineering) 76
- データ型 (data type) 107
- データ行列 (data matrix) 123
- データ行列の分散 (variance of data matrix)
 - 127, **146**
- データサイエンス (data science) **75**, 77
- データサイエンティスト (data scientist)
 - 70, **95**
- データテーブル (data table) **9**, **102**
- データ同化 (data assimilation) 77
- データフレーム (data frame) .. **102**, 123, 135
- データブローカー (data broker) 119
- データ分析 (data assay) 7
- データ分布 (data distribution) 27
 - の位置 (location) 35
 - の代表値 (summary) 35
 - の中心 (center) 35
 - の広がり (dispersion) 35
- データベクトル (data vector) **102**, 124
 - が直交 (orthogonal) 201
- データベンダー (data vender) 119
- データマイニング (data mining) 78
- データリテラシー (data literacy) 89, 96
- テーブル (table) **8**
 - の結合 (join) 102
 - の差分 (difference) 101
 - の射影 (projection) 102
 - の選択 (selection) 102
 - の直積 (direct product) 101
 - の併合 (union) 101
- TextilePlot 157
- テキストファイル (text file) 98
- テキストマイニング (text mining) 197
- デシベル (decibel) 45
- 天井関数 (ceiling function) 38
- 転置 (transpose) 101
- 転置ベクトル (transposed vector) 124
- デンドログラム (dendrogram) 135

258 索引

と

統計的仮説検定 (statistical hypothesis testing) 223
 同時確率分布 (joint probability distribution) 209
 同時確率分布関数 (joint probability distribution function) 209
 等質 (homogeneous) 5
 等質性 (homogeneity) 83
 同時分布 (joint distribution) 64
 東証株価指数 (Tokyo stock price index) 237
 淘汰 (selection) 132
 同定 (identify) 103
 特異値 (singular value) 154
 特異値分解 (singular value decomposition) 153
 特異ベクトル (singular vector)
 左— 155
 右— 155
 独立 (independent) 206
 度数 (frequency) 29
 度数分布多角形 (frequency polygon) 32
 度数分布表 (frequency table) 29
 ドメイン (domain) 9, 107
 トレース (trace) 127

な

生データ (raw data) 11, 48, 79

に

2 元度数分布表 (two way frequency table) 63
 2 項確率 (binomial probability) 230
 2 項係数 (binomial coefficient) 230
 2 項分布 (binomial distribution) 91
 2 次形式 (quadratic form) 145
 2 次元散布図 (two dimensional scatter plot) 16
 2 次データ (secondary data) 48, 83, 108
 日経平均 (Nikkei 225) 56, 237

ニューラルネットワーク (neural network) 78

の

ノイズ (noise) 211
 ノット (knot) 162

は

パーセント点 (percentile) 206
 バイナリーファイル (binary file) 100
 バイプロット (biplot) 156
 配列 (array) 105, 123
 箱型図 (boxplot) 51
 箱ひげ図 (box whisker plot) 51, 89
 ハザードレート (hazard rate) 233
 外れ値 (outlier) 44, 53, 93
 バラツキ (variability) 145
 パラメータ (parameter) 6
 範囲 (range) 21
 範囲幅 (spread) 20
 判別分析 (discriminant analysis) 160
 凡例 (legend) 87

ひ

ピアソンの相関係数 (Pearson's product-moment correlation coefficient) 205
 BI (business intelligence) 77, 120
 PCA (principal component analysis) 142
 p 値 (p -value) 97, 183, 206, 223
 BPT モデル (Brownian passage time model) 244
 P-P プロット (probability-probability plot) 222
 ピクセル (pixel) 45
 ヒストグラム (histogram) 29, 88
 ヒストリカルデータ (historical data) 118
 非斉次ポアソン過程 (inhomogeneous Poisson process) 232
 被説明変量 (explained variate) 104, 167

ピタゴラスの定理 (Pythagorean theorem) **24, 127**

ビッグデータ (big data) **3, 70, 71, 121**

秘匿 (conceal) **36**

標準偏差 (standard deviation) .. **21, 41, 207**
 度数分布表から求めた— **42**

標本相関係数 (sample correlation coefficient) **199**
 頑健な (robust)— **205**

標本相関係数行列 (sample correlation coefficient matrix) **200**

標本調査 (sampling survey) **111**

標本分散共分散行列 (sample variance covariance matrix) **200**

標本偏相関係数 (sample partial correlation coefficient) **199, 202**

標本偏相関係数行列 (sample partial correlation coefficient matrix) **202**

ふ

複式簿記 (double-entry bookkeeping) **13**

符号関数 (sign function) **40**

不偏分散 (unbiased variance) **23**

付与属性 (given attribute) **106**

ブラウザ (browser) **100**

ブラウジング (browsing) **71, 80**

ブラウン運動 (Brownian motion) **243**

フラクタル (fractal) **218**

フラクタル次元 (fractal dimension) **219**

フラットファイル (flat file) **98**

プレゼンテーション (presentation) **87**

分割表 (contingency table) **193**

分散 (variance) **21, 207**
 度数分布表から求めた— **42**

分散共分散行列 (variance covariance matrix) **127, 208**

分散分析 (analysis of variance) **182**

分布 (distribution) **27**

分布関数 (distribution function) **206**

分類変量 (classification variate) **104**

へ

平滑化 (smoothing) **187**

平滑曲線 (smooth curve) **90, 187**

平均絶対偏差 (mean absolute deviance) ... **42**

平均値 (mean, average) **18, 35**
 度数分布表から求めた— **36**

並行座標軸プロット (parallel coordinate plot) **157**

ベイジアンネットワーク (Bayesian network) **78, 214**

ベイズ統計学 (Bayesian statistics) **77**

ベイズの公式 (Bayes formula) **214**

ヘッダー (header) **99**

別名 (alias) **84**

ヘビーテール (heavy tail) **50**

変異 (mutation) **132**

偏共分散 (partial covariance) **210**

偏差 (deviation) **21, 41**

偏差値 (standard score) **24**

偏差ベクトル (deviance vector) **125**

変数 (variable) **5**

偏相関係数 (partial correlation) **210**

変動係数 (coefficient of variation) **44**

変容 (transfiguration) **20**

変量 (variate) **4, 101**
 —の型 (type) **12**

変量重み行列 (variate weight matrix) ... **145**

変量空間 (variate space) **123, 125, 144**

ほ

ポアソン確率 (Poisson probability) **229**

ポアソン過程 (Poisson process) **229**

ポアソン現象 (Poisson phenomenon) **230**

- ロジット (logit) 184
- ロジットモデル (logit model) **184, 185**
- ロバスト回帰 (robust regression) 172
- ロバスト推測 (robust inference) 94
- ロングテール (long tail) 49