はしがき

本書は，データ分析に必要な統計の基礎を述べている．統計を用いたデータ分析は，現在ではさまざまな分野で行われている．従来から工学，心理学，経営学，医学，経済学，政治学，文学，考古学，社会学などでデータ分析は行われてきた．さらに，音楽や文学作品から収集したデータの分析も行われている．このように，データ分析を行う分野が拡がるとともに，データを収集する手段も多様になった．例えば，調査であれば，従来は調査票を郵送し，調査対象者に記入後返送してもらったり，あるいは，調査員が調査対象者を訪問して調査票に記入して調査票を回収するという手順が一般的であった．しかし，現在では，ウェブ・サイトを利用した調査なども行われている．もちろんこのようなデータ収集法にまったく問題がないわけではないが，従来に比べて遥かに安価な費用でデータを収集することができるようになった．また，百貨店やスーパーマーケット，あるいは，レンタルショップの顧客カード，IC 乗車カード（スイカやイコカなど），銀行のキャッシュカードなどでは，使用者の移動履歴，購買履歴，閲覧履歴などのデータが容易に収集できる．カードなどの所有者が特定されている場合には，所有者の情報（年齢，性別，住所，その他）と移動履歴，購買履歴，閲覧履歴などを組み合わせて分析することで，データ分析を通じてさまざまな情報が得られる．さらに，ウェブ・サイト経由の通信販売の履歴，ウェブ・サイトの閲覧履歴などを分析することも広く行われている．したがって，データを収集するさまざまな手段が用いられるようになったことで，データは量的には増加し，質的には多様になったのである．

このようなデータを分析するためには，多くの場合に統計的な方法を用いる．データ分析を行うためには，既成の統計パッケージや R 言語などを用いることになる．これらを利用する際には，データなどの入力の形式さえ適合していれば，容易に分析結果が得られる．既成の統計パッケージや R 言語などから分析結果が得られることと，データ分析を適切に行うことは別である．これらをデータ分析で正しく利用し，適切にデータ分析を行うためには，データ分析で利用される統計の基礎的な知識が欠かせない．本書は，このようなデータ分析に必要な統計の基礎知識を述べた入門書である．

データ分析は，何らかの判断や意思決定における結論を導くために行うことが多い．統計を用いたデータ分析によって得られる結果は，数量的に表されているという意味で，また，誰が分析してもデータが同じであれば同じ結果が得られるという意味で客観的である．さらに，結果を導く過程が論理的であるという点も統計を用いたデータ分析の特徴である．しかし，結果が論理的に導かれ，数量的に表され，客観的であることから，統計を用いたデータ分析によって得られる結果を必要以上に過大評価し，得られた結果がそのまま判断や意思決定における結論そのものであると考えるべきではない．なぜならば，判断や意思決定のために必要な情報のすべてがデータとして収集されているわけではなく，統計を用いたデータ分析で扱ったデータは，このような情報の一部に過ぎない．また，データ分析で利用される統計手法にもそれぞれの制約があり，データのもつ情報の一部だけ

しか扱うことができない．したがって，データ分析から得られた結果は，何らかの結論を求めるための判断や意思決定のための 1 つの資料であると考えるべきである．実際に判断や意思決定において何らかの結論を得るためには，判断する人あるいは意思決定する人が，データ分析で得られた結果とデータ分析で扱われなかった情報，さらには，経験などを総合的に考えて行わなければならない．他方，データの一部だけを収集し，そのデータのもつ情報の一部だけを数量的に分析するということから，得られた結果は実際の判断や意思決定では役に立たないのだと考え，経験や直感だけに頼って主観的また定性的に結論を導くのも好ましくない．データ分析で得られた結果は，意思決定や判断で結論を導くための資料であり，提供された資料は活用すべきなのである．

本書の特色は，統計の理論的な側面よりも，考え方や概念，背景にある原理，これらのもつ意味の記述に重点をおいたことである．そして，これらが視覚的に目で見て理解できるように，図を多く用いて記述を進めたことも本書の特色である．本書の草稿は，岡太が 1 章，2 章，3 章の 3.1.1 項，3.1.2 項，3.2.1 項，3.2.2 項，3.2.4 項の一部，4 章，5 章，6 章，7 章，8 章，および，9 章を担当し，中井が 3 章の 3.1.3 項と 3.2.3 項，3.2.4 項の一部，および，10 章を担当し，元治が 11 章を担当した．草稿をもとに 3 人の執筆者がさまざまな面から検討し，推敲を重ねて草稿を改善し，最後に岡太が全体的な調整を行った．本書に示した計算は，実際には表記した桁数よりも多い桁数で行っている．本文に表記された桁数で計算を行うと実際の計算よりも精度が低下するため，本文に書かれている計算結果の値と多少の差異が生じる場合がある．例えば，例 8.7 (p.90) の下限の計算では，下限 $= 0.28 - 1.9600\sqrt{0.28\,(1-0.28)/254} = 0.23$ である．この計算では，0.28 でなくデータから計算した比率 72/254 が用いられている．0.28 を用いて計算すれば 0.22 (0.22478) となる．

さまざまな事情で当初の予定よりも遅れてしまったが，共立出版株式会社の吉村修司氏のひとかたならぬお骨折りにより，出版にいたることができた．ここに記して心よりの感謝を表す次第である．吉野諒三氏（統計数理研究所教授）には RDD と呼ばれる電話を用いた調査についてご助言を頂戴した．山口和範氏（立教大学経営学部教授）は，11 章を読んでくださり貴重なご助言を頂いた．横山暁氏（帝京大学 経済学部助教）には，草稿を読んでわかりにくい表現や誤解しやすい記述をご指摘頂き，また章末問題を実際に解答くださり，多くの有益なご意見を頂いた．長尾二郎氏（青木・関根・田中法律事務所弁護士）には，例などの引用についてご丁寧なご助言を頂いた．ここに記して深い謝意を表す．

2012 年 9 月
岡太 彬訓