探検データサイエンス　データサイエンスのための確率統計
尾畑 伸明著
https://www.kyoritsu-pub.co.jp/bookdetail/9784320125186

まえがき

本書の目的は，データ科学（データサイエンス）の数理的基礎付けとして確率と統計を学ぶことである．データ駆動型社会の到来にあたって，確率と統計の基礎知識はデータを実務として扱うことがなくとも，日常生活の中で様々に現れるデータを的確に理解し役立てるためにもますます重要になっている．近年，高等学校においても確率と統計を学ぶ機会が増えている．本書は高等学校で学ぶ確率のごく初歩と数学リテラシーを前提とするが，必要に応じて補習すれば十分であろう．微積分の数式が時々出てくるが，本書のレベルでは微積分の計算はあまり必要ない．数式を見た方が理解が早いという読者には好都合であるし，数式が苦手であってもそれを概念的にとらえることで十分に役に立つ．最後の章（第 12 章）では多変量データを扱うが，そのために必要となる「行列」については必要最低限の記述で済ませた．多変量データ解析を志す読者は，行列について線形代数の教科書などで一通り学ぶことをお薦めしたい．

統計学は伝統的に，収集されたデータから有用な情報を抽出して，データの意味や価値を見出し，合理的な意思決定を行うことを目的として体系化された学問である．可能な限り大量の資料を収集して，そこから何らかの情報や法則を得ようとする試みは，国家の現況を把握するための人口統計を例に見ても実に古くから行われてきた．しかし，数理的な根拠をもって統計を扱い始めたのは 19 世紀以降であり，推測統計学として今日の形式が整ったのは 20 世紀になってからである．その基本的な立場は，収集されたデータは調査対象（母集団）の一部にすぎないので，母集団の性質を完全に復元することはできず，あくまで推測というところにある．推測に合理性をもたせるためには信頼性の評価が欠かせないが，20 世紀初頭に確立した現代確率論という強力な数学を用いることで，推測統計学として発展して今に至っている．

ところが，今日の状況は 20 世紀初頭とは大きく異なる．情報通信技術や計算機の発展とともに，計算機の利用を前提とした大規模なデータ（ビッグデー

探検データサイエンス　データサイエンスのための確率統計
尾畑 伸明著
https://www.kyoritsu-pub.co.jp/bookdetail/9784320125186

タ）が身近になったのである．初期の推測統計学では少数のデータを大切に扱って，できるだけ正確な推測をすることが重視されていたといえる．その観点の重要性は，少数データを扱う限り今日でも変わらない．一方，大規模データの解析では，データから法則性を見出し，構造的な特徴を抽出し，それを説明するモデルの導入（モデリング）がより重要になる．モデリングそのものは伝統的な統計学においても主要な課題の一つであるが，従来の手法に加えて機械学習に代表される強力な計算技術が開発され，データ科学の数理として大きく展開する時代の到来が感じられる．

本書は確率論の基礎から始めて，推測統計の典型的なテーマである母数の推定と検定を扱う．統計処理の処方箋の提示で終わることなく，理論的な仕組みをできるだけ丁寧に記述することで，より深い理解を期待している．さらに，多変量解析の典型として回帰分析と主成分分析の基本を学ぶ．データを説明するモデルの構築として線形回帰モデルを扱い，構造的な特徴の抽出として主成分というアイデアを紹介する．例題のデータ数は大きくはないが，データの配置や統計量の計算を実感するために表計算ソフトが使えると助けになる．もちろん，Python などのプログラミング言語が使えれば，いろいろな試みが可能になる．

本書を「探検データサイエンス」シリーズの 1 冊として執筆できたことに対して関係諸氏に謝意を表したい．本書がデータサイエンスを志す方々にとって少しでも役に立てれば著者にとって誠に幸いである．最後に，著者の遅筆に関わらず，励まし続けてくださった共立出版編集部の山内千尋氏と天田友理氏に感謝したい．

2021 年 6 月

尾畑伸明