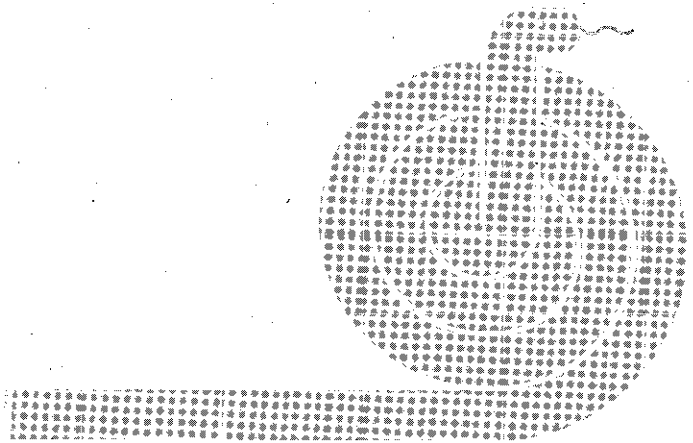
いまさら聞けない

Pythonでデータ分析

多変量解析，ベイズ統計分析（PyStan，PyMC）

OKAMOTO YASUHARU
岡本安晴



丸善出版

はじめに

本書は，データ分析を Python で行うときの基礎的事項の解説を試みるものである．Python は，プログラミング言語としてはやさしい使い方もできるが，いろいろなライブラリ・パッケージが用意されているので，様々な分野で用いられている汎用性の高い言語でもある．

読者として，初めて Python でデータ分析を行ってみようと思う人を想定している．Python は全く初めてという人も本書で Python が使えるように工夫したが，言語の文法の詳しい解説は他書に譲った．データ分析に関わる基礎的事項はできるだけ扱うようにして，本書を辞書的に参照して長く利用されるようにした．

まず，データおよびその分析結果の視覚化として Python のライブラリ matplotlib を用いたグラフ描画について説明するが，これは本書全体にわたる分析の視覚化において用いられる．データ分析法としては，基本となる標準的な多変量解析と現在注目を集めている確率モデルによるベイズ分析を取り上げる．多変量解析においては行列演算を用いるのが現在の標準的方法であるが，ライブラリ numpy を用いた行列演算について説明する．ベイズ分析については，Stan の Python 用である PyStan，および Python 専用の PyMC を取り上げる．いずれも，基礎的使い方の説明を心がけた．

各章の内容は以下のとおりである．

第１章　基本統計量の計算

平均，分散など基本統計量を取り上げて，Python のコード例を示した．Python プログラミングの雰囲気を示すとともに，numpy などのライブラリの利用例も示した．

第２章　グラフ描画――データの可視化――

情報は視覚化されるとわかりやすい．データおよびその分析結果の視覚化はグラフという形で表示される．Python でのグラフ描画のライブラリとして matplotlib を取り上げる．棒グラフ，ヒストグラム，折れ線グラフを例に挙げて，その使いやすさを示す．

第３章　ファイル入出力

簡単なデータは，Python スクリプト中に書き込むことができるが，実用的にはファイルから，あるいは適当なストリームからの入力になる．本章において，ファイル入出力の方法として，テキストファイル，CSV (Comma-Separated Values)形式ファイル，バイナリファイルを取り上げる．

第 4 章　行列演算と Python スクリプト

多変量解析の基礎である行列演算を取り上げる．Python における行列演算のライブラリとして numpy を取り上げ，行列演算が簡単に行えることを説明する．

第 5 章　単回帰分析

変数間の関係を 1 次式で表すのは，統計分析の 1 つの出発点である．その最も簡単なモデルが 2 変数の関係を 1 次式で表す単回帰モデルである．この単回帰モデルを行列演算で表し，その解が行列で簡単に表せることを説明するとともに，numpy の ndarray を用いたスクリプト例を示す．

第 6 章　重回帰分析

重回帰分析では，複数の独立変数の影響が 1 次式で表される．行列を用いると，形式的には単回帰モデルと同様に扱える．しかし，複数の独立変数を用いることにより，単回帰モデルでは表されない関係を扱うことができるが，これは演習課題とした．

第 7 章　主成分分析

多くの変数から構成されるデータは，その主な情報がいくつかの成分で表されることが多い．この成分を変数の 1 次式で求める方法として主成分分析があるが，本書ではこれを正射影の観点から説明した．主成分分析も行列を用いると簡単に表すことができる．

第 8 章　数量化

調査あるいは質問紙データは，カテゴリ変数がよく用いられる．カテゴリ変数を数量化すると，他の数量変数とともに標準的な多変量解析を適用することができる．数量化の計算は，行列を使えば簡単である．数量化のための Python スクリプト例を示す．

第 9 章　確率計算と Python スクリプト

ベイズ分析は，確率モデルの強力な方法として注目を集めている．確率計算の Python スクリプト例とともに，ベイズ分析の基本モデルの簡明さを示す．

第 10 章　PyStan による 2 項分布分析—— Stan 入門——

ベイズ分析は，事後分布をシミュレーションで求める方法が開発されて実用性と有効性が飛躍的に高まった．Stan は，シミュレーションで求めるライブラリの 1 つであり，Python 用の Stan が PyStan である．基礎的な使い方は簡単であり，簡単な確率モデルである 2 項分布を使って説明する．

第 11 章　PyStan による単回帰モデル分析

統計分析の基本モデルとして単回帰モデルを取り上げ，PyStan によるベイズ分析について説明する．

第 12 章　PyStan によるポアッソン回帰モデル分析

回帰モデルの一般化の 1 つとして，ポアッソン回帰モデルを取り上げる．また，分散分析のように要因がカテゴリ変数のときは，ダミー変数を用いると回帰モデルが設定できる．クロス表の分析を，ポアッソン回帰モデルにおいてダミー変数を設定して行う．

第 13 章　PyMC による 2 項分布分析—— PyMC 入門——

PyMC によっても，ベイズ分析における事後分布をシミュレーションで求めることができる．基本的方法は簡単であり，簡単な確率分布として 2 項分布モデルを取り上げて説明する．

第 14 章　PyMC による単回帰モデル分析

統計分析の基本モデルの 1 つとしての単回帰モデルを取り上げ，PyMC による分析について説明する．

第 15 章　PyMC によるポアッソン回帰モデル分析

回帰モデルの一般化の 1 つとして，ポアッソン回帰モデルを取り上げる．また，分散分析のように要因がカテゴリ変数のときは，ダミー変数を用いると回帰モデルが設定できる．クロス表の分析を，ポアッソン回帰モデルにおいてダミー変数を設定して行う．

本書の企画段階から 1 次原稿まで，いろいろな方々から御意見，御提案，励ましをいただいた．ここにすべての方のお名前を挙げることはできないが，謝意を表する次第である．特に大阪大学の足立浩平教授および狩野裕教授の研究室の方々からコメントやアドバイスなどをいただいたことは記しておきたい．また，加藤直哉氏と加藤仁美氏には，初校前の原稿について励みとなるコメントをいただいたことを記しておきたい．もちろん，本書に残る誤りなどの問題点は，最終的に著者の責任であることは言うまでもない．

丸善出版株式会社企画・編集部第三部長の小西孝幸氏には，筆者の最初の企画提案の段階から有益な助言をいただいた．本書の現在の内容は，氏の援助に基づくところが大きい．著者は当初，データ分析・統計分析の分野に進もうという学生でプログラミングが初めてという人を対象としたPythonの入門書を考えていた．小西氏からPython関連書籍の情報などが提供され，著者の研究および授業におけるプログラミングの経験をよりよく反映した本書を上梓することができた．ここに記して謝意としたい．

2018年　夏

自宅にて
岡本安晴

目　次