まえがき

本書はPythonによるデータ分析の書籍を読むための入門書です。話題のPythonを使ってデータを分析してみたい。ただ，プログラミング経験がまったくないので不安だ，もしくはPythonの入門書を読み始めたけど挫折したという読者が対象です。

本書はこれまでのPython関連書籍を補完します。文系の読者が挫折しやすいのは，理系向けの書籍は統計学・数学の説明が中心だからです。一方で，Pythonの使い方を簡単に解説した書籍が物足りないのは，どんな場面に応用できるかを踏み込んで示していないからです。

そこで，本書では細かいことから積み上げず，こんなことができるという事例を概観して，イメージを持つことを重視します。このため，説明は短く，数式や専門用語はほぼなし，各章ごとのテーマを体験しながら学ぶことが特徴になっています。

途中で挫折しない工夫もしています。各章は独立しており，経営学，経済学，社会学に関連した事例を取り上げ，興味のある章から読めるように配慮しています。

Pythonや統計学・数学を1からすべて学ぶのは大変です。興味のあるトピックスに関連する部分から学べばよいのです。その際，Pythonの使い方だけではなく，分析手法の応用の仕方を学びましょう。

また，Pythonの関連書籍で初学者がわからなくなりやすい部分や警告・エラーが出る場合にどうしたらよいかや，さらにちょっと深く理解したい場合の補足などを「よくある質問」「たまにある質問」というコラム形式で回答しています。

いろいろ詰め込むのではなく，本文は骨子にとどめ，補足を「質問」にしたわけです。これは理解を助ける工夫です。コンパクトにすることで読みやすくしただけでなく，質問形式だと頭に入りやすいというメリ

ットがあります。とにかく手っ取り早くという要望に応え，できるだけそぎ落として読みやすくしたのが特徴です。

■個人で使用される方へ

とりあえず，本書の内容を写経のようにとにかく打ち込んで慣れることから始めましょう。はじめのうちはエラーが多く，思いのほか時間がかかるかもしれません。ただ，続けるのが大事です。だんだんと慣れてきます。

■講義で使用される方へ

1章1回分，半期を目安にされるとよいでしょう。本当に右も左もわからない人向けの内容なので，社会科学にとどまらず，広くいろいろな分野の初心者向けの講義（高校生の授業を含む）にも適していると思っています。

本書が，将来，次のステップの勉強に進むための役に立てれば幸いです。その意味で，Python によるデータ分析の書籍を読むための入門書としています。このため，厳密な説明や用語というよりは，わかりやすさを優先したところがあることをご理解いただけると幸いです。

最後になりましたが，本書のタイトルを含めいろいろご提案いただいたり，新しいバージョンのコードに修正していただいたりした有斐閣の渡部一樹さん・渡辺晃さんに感謝申し上げます。本書がわかりやすい内容になっているとすれば丁寧なコメントをくださったおふたりのおかげです。

2024年9月

友原 章典

動作環境

本書で示されるプログラムは，次の環境で動作確認しています（2024年7月時点）。

OS：Windows 11 Home バージョン 22H2
Python：Python 3.11.8
Jupyter-notebook：7.0.6
Web ブラウザ：Google Chrome

beautifulsoup4	4.12.2	pandas	2.1.4
ipykernel	6.28.0	pip	23.3.1
ipython	8.20.0	PuLP	2.7.0
matplotlib	3.8.0	scikit-learn	1.3.0
mecab-python3	1.0.8	seaborn	0.12.2
networkx	3.1	wasabi	0.10.1
numpy	1.24.3	watchdog	2.1.6
openpyxl	3.0.10	wordcloud	1.9.3

注意事項

さまざまな要因が影響しうるため，上記の条件を再現しても本書で示されるプログラムの実行結果がすべての環境で再現されるとは限りません。本書の内容は 2024 年 3 月の原稿執筆時点のものであり，本書に掲載したソフトウェアのバージョンや URL，プログラムの出力結果，操作方法・手順などは変更される可能性があります。

プログラムファイルの作成にあたっては，内容に誤りのないようできる限り注意を払いましたが，結果生じたこと（損害等）には，著者，出版社とも責任を負うことはできません。

本書に記載されている会社名，製品名ならびにサービス名等は，各社の商標および登録商標です。

目 次

＊……Python の文法やデータ分析について，やや発展的な話題を扱っていることを意味します。はじめは読み飛ばしてもかまいません。

ウェブサポート

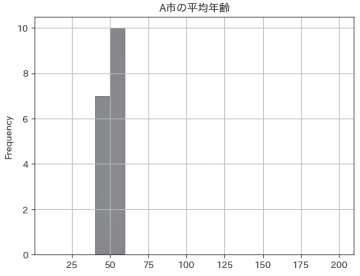
本書で使用するデータや練習問題の解答例など，各種サポート情報を下記のページから提供していきます。ぜひご利用ください。

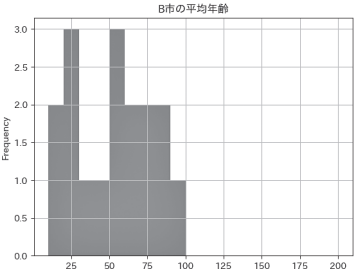
https://www.yuhikaku.co.jp/books/detail/9784641166363

第3章

特徴を踏まえて適切な計画を立てよう

平均とヒストグラム





市町村の住民の特徴を比べて，地域住民の違いを調べます。住民の特徴を1つの数値で示して比べることで，適切な政策を考えてみましょう。たとえば，住民の平均年齢が50歳と同じであるA市とB市を例にとります。平均年齢で見た住民の特徴が同じなので，同じ政策（例：年金生活者に配慮する政策）を行えば，いずれの市の住民も満足するでしょうか。

1 平 均

1 データの入力と住民数の表示

A 市，B 市，C 市という 3 つの市の住民の年齢について架空データを作りましょう（C 市には 179 歳以上の年齢が入っていますが，データ分析の練習のためのデータなので気にせず入力してください）。最初の 4 行については後ほど説明します。

コード 3-1

```
import pandas as pd
import matplotlib.pyplot as plt
import japanize_matplotlib
import seaborn as sns

data = {
    "A 市":[
        42, 43, 45, 46, 47, 48, 49, 50, 51,
        52, 52, 52, 52, 53, 55, 56, 57,
    ],
    "B 市":[
        10, 14, 20, 23, 28, 33, 40, 51, 53,
        54, 62, 62, 70, 75, 80, 85, 90,
    ],
    "C 市":[
        8, 13, 18, 26, 27, 29, 33, 35, 37,
        38, 39, 45, 45, 48, 179, 186, 198,
    ],
}

df = pd.DataFrame(data)
df.head()
```

↳出力結果

	A 市	B 市	C 市
0	42	10	8

1	43	14	13
2	45	20	18
3	46	23	26
4	47	28	27

df.head() で最初の 5 行（それぞれの市に住む 5 人の年齢）が示されます。前章で述べたように，Python では最初の行は「0」から始まります。

次にデータの「長さ」を調べる len() を使って，住民数を示します。

コード 3-2
```
len(df)
```

出力結果
```
17
```

それぞれの市に住民が 17 人いることがわかります。

ここではすべての市で住民数が同じですが，A 市の住民数だけを示したいときには次のように入力します。

コード 3-3
```
len(df["A 市"])
```

出力結果
```
17
```

2 平均の計算

mean() を使って，それぞれの市の年齢の**平均値**（mean）を計算します。これまではセルの最終的な処理結果をそのまま出力することが多くありましたが，出力をより見やすくするために，ここでは print() を使います。出力したい内容をカンマ区切りで指定すると，それぞれをスペー

ス区切りで出力してくれます。

コード3-4
```
print("A市", df["A市"].mean())
print("B市", df["B市"].mean())
print("C市", df["C市"].mean())
```

出力結果

```
A市  50.0
B市  50.0
C市  59.05882352941177
```

なお，df.mean()と入力すると，市ごとに計算するのではなく，すべての市の平均年齢を一度に計算できます。

コード3-5
```
df.mean()
```

出力結果

```
A市  50.000000
B市  50.000000
C市  59.058824
```

A市の平均年齢は50歳。B市の平均年齢も50歳です。C市の平均年齢は約59歳です。

2 ヒストグラムで分布を把握する

A市とB市の平均年齢は同じですが，データを見るかぎり，2つの市の特徴が同じようには見えません。そこで，年齢構成の特徴を可視化するために，グラフを描いて，詳しく見てみましょう。

1 ヒストグラムの描き方

新たなライブラリを使って，A 市の住民の年齢の分布を図で描いてみましょう。分布図のことを**ヒストグラム**（histogram）といいます。

コード 3-6

```
import pandas as pd
import matplotlib.pyplot as plt
import japanize_matplotlib
import seaborn as sns

age_spans = [
    10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 110,
    120, 130, 140, 150, 160, 170, 180, 190, 200,
]

df["A 市"].plot(
    kind="hist",
    bins=age_spans,
    title="A 市の平均年齢",
    grid=True
)
```

最初の 4 行の import で，計算したり，図を描いたりする際に必要なものをインポートします。

pandas は第 2 章でも使いましたが，データの取り込みや分析などに使うものです。matplotlib.pyplot は図を作成するときに使うもので，japanese_matplotlib は日本語で表記されている変数名（売上高など）が図にきちんと表示されるようにするためのものです。seaborn は図をより見やすくするために使います（本章中のコードでは使用しませんが，練習問題で使用します。これら 4 つはセットで使用することが多いので，「おまじない」のようにノートブックの先頭に書くクセをつけておくのもよいでしょう）。

ここでは，as を使うことで，pandas を pd，matplotlib.pyplot を plt などと省略して使えるようにしています。これによってコードを書く負担が減ります。

よくある質問

import はまとめて記述する必要があるのですか？

Jupyter Notebook では，一度ライブラリをインポートすれば，ノートブックのそれ以降の箇所でいつでも使えるので，まとめて書く必要はありません。ただ，使用するライブラリをまとめて書くことで，そのノートブックで行う分析の概要を示せてわかりやすいので，まとめて書く慣習があります。

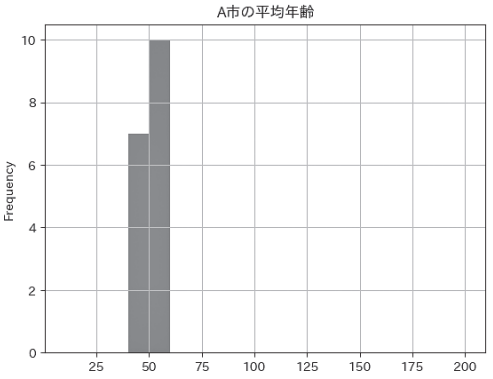
仕事（＝分析）をするにあたり，能力（＝機能）に応じて，一緒に働く仲間を招集するイメージです。今回は，4 行なので 4 人のパーティメンバーを集めたといったところでしょうか。

age_spans の部分でたくさんの数字を書いているのは，年齢幅を 10 歳区切りにするためです。

plot() の各引数を指定することでヒストグラムを生成できます。kind は図の種類をヒストグラム（hist）として指定するもので，bins ではデータの区切りを指定しています。title と grid の引数はその名のとおり，タイトルとグリッド（目盛り線）を指定するオプションです。

図 3-1 のようなグラフが出力されます。

図 3-1　コード 3-6 の出力結果



横軸が年齢，縦軸は住民数です。40〜49 歳の住民が 7 人，50〜59 歳の住民が 10 人いるのがわかります。

A 市では，住民の年齢が 50 歳付近に集中しています。

たまにある質問

図の日本語が文字化けします。どうすればいいですか？

japanize_matplotlib が正しくインストールされていない可能性があります。エラーが表示されていたらその文面でネット検索することで解決のヒントが得られます。

どうしてもエラーが解消しないときには，sns.set(font="Yu Gothic") や sns.set(font="Meiryo") のように個別にフォントを指定しても大丈夫です。それぞれ游ゴシックやメイリオのことです（Jupyter Notebook を実行しているパソコンにインストールされているフォントでないものを指定しても反映されないのでご注意ください）。

import japanize_matplotlib の 1 行を入れると，sns.set() を使う必要はありません。どちらか 1 つで構いません。

B 市の住民の年齢の分布を図で描いてみましょう。

コード 3-7

```
df["B 市"].plot(
    kind="hist",
    bins=age_spans,
    title="B 市の平均年齢",
    grid=True
)
```

図 3-2 のようなグラフが出力されます。

B 市では，住民の年齢が 10〜100 歳の間に散らばっているのがわかります。

著者紹介　友原 章典（ともはら・あきのり）
青山学院大学国際政治経済学部教授
早稲田大学政治経済学部卒。ジョンズ・ホプキンス大学大学院Ph.D.（経済学）。米州開発銀行，世界銀行コンサルタントから，ニューヨーク市立大学助教授，UCLA 経営大学院エコノミストなどを経て現職。『移民の経済学――雇用，経済成長から治安まで，日本は変わるか』（2020 年，中央公論新社），『会社ではネガティブな人を活かしなさい』（2021 年，集英社）など。

文系のための Python データ分析――最短で基本をマスター

Introduction to Data Analysis with Python

2024 年 10 月 30 日 初版第 1 刷発行

著　者　友原章典
発行者　江草貞治
発行所　株式会社有斐閣
〒101-0051 東京都千代田区神田神保町 2-17
https://www.yuhikaku.co.jp/
装　丁　嶋田典彦（PAPER）
印　刷　大日本法令印刷株式会社
製　本　牧製本印刷株式会社
装丁印刷　株式会社亨有堂印刷所

落丁・乱丁本はお取替えいたします。定価はカバーに表示してあります。

Printed in Japan. ISBN 978-4-641-16636-3