第一回：統計学の基礎

尚　晋
大学院経済学研究科助教

2025年4月15日

今日のポイント

1. 統計学とは
2. 統計的推測とは
3. データの種類、集め方と測り方

1　統計学とは

1.1　いくつかの定義

定義：統計学
　データを分析し、有用な情報を取り出す方法論のこと。
定義：記述統計学
　データを整理・要約すること。
定義：統計的推測
　一部の観察から全体について推測すること。

1.2　統計学の目的

統計学 statistics は今でこそ整然とした理論体系をなしているが，もともとは，人間やその社会におけるさまざまな実践的関心や活動から起り，その考え方や蓄積した知識が合流して太い流れをなしたものである.例えば、

- ゲームのテーブルから起った確率論
- 常備軍や国家財政上の必要から起った国家状態の統計
- 17 世紀のペスト禍を機とする近代死亡率表の研究
- 生物等で生じる諸量の相関関係の理論
- 経済学や気象学における時系列の理論

などが挙げられる．

　このことからわかるように，現象の法則性に対する人間のあくなき実際的関心が統計学を生み出した．これが「統計学とは何か」という問題に対する手短かな答えである．現象の法則性を知るために，データを丹念に調べ，規則性から法則を見出してもよいし，また，一部を観察して，そこから論理性のある推測で全体の法則性の発見にいたってもよい．

補足 1.1 全体（集団）の法則性を知るには全体を調べねばならない，ということは理屈の上ではもっともで，これが全数調査といわれる考え方である．国勢調査censusは忠実にこの原理を実行している．

　Statistik(英statistics)とは，元来は今日の統計学をさすものではなく，国家(Staat, [英]state)の状況を歴史的に記述する，端的にいえば今でいう統計と呼んでいいものである．

　「政治算術」や「国勢学」とは違って，今日の統計学は，科学的推論のための方法論の体系という特色をもっている．この方法論は，確率論に基づく数学的根拠とともに，広い応用範囲と妥当性をもっている．

　大量観察は，ある一定の普遍性をもった法則をもたらし，その法則性を確定するのが，統計学の役割であると考えられている．

　標本sampleから母集団populationへと，法則が存在する所がはっきりとし，二つが区別して考えられ，そこに確率論が本格的に用いられることになった．これを基礎に，われわれが今日用いている統計的推測statistical inferenceの論理が築かれた．母集団に対する推定estimationの理論と仮説検定hypothesis testingである．

　母集団全体を調べるのは物理的・制度的に難しい。母集団から標本を抽出⇒標本を分析して、母集団の様子を推定

　例：日本の会社員全体（母集団）の平均年収を調べるのは難しい⇒全国から無作為に抽出された 1000 人の会社員（標本）の平均年収を計算。

　確率論：不確実なことがらを、数理的に処理するための道具。限られた標本による分析結果は、誤差・不確実性を伴う⇒　確率論を応用。

2　統計データと統計手法

2.1　データの用語

◆データの例：高校生の身長・体重・性別

番号	身長（cm）	体重（kg）	性別（女=1）
1	178	63	0
2	165	62	0
3	168	69	0
4	152	41	1
…			
15	168	59	1

◆データの用語

　観測個体：記録された個体ひとつひとつを観測個体と呼び、代表してiと表す。例：i = 2（2 番目）の個体は、身長 165cm、体重 62kg、男性

　変数：記録されている個体の情報。例：このデータの変数は、身長、体重、性別の三つ

　次元：変数の数。例：このデータは三次元のデータ

　サンプル数：観測個体の総数とサンプル数（或いはサンプルサイズ）呼び、nで表す。例：上のデータのサンプル数はn = 15

2.2　データのタイプ

◆量的データと質的データ

量的データ：長さや重さ、金額、温度、時間など。定量的に測られたデータのこと。例：身長と体重。

質的データ：性別（男・女）や学歴（中卒・高卒・大卒）など、個体の属性・状態を示すデータのこと。例：性別。

質的データは，ダミー変数と呼ばれる0または1をとる変数によって数量化が可能となる（たとえば，性別の場合は男性を0、女性を1とする）

◆1次元データと多次元データ

1次元データ：一人の学生に対して一つの観測値（身長）だけが与えられる場合、このようなデータを1次元データ(1-dimensional data)と呼ぶ．1次元データに関しては，度数分布表を描いたり平均などの代表値や分散を求めて分析を行う．

多次元データ：一人の学生に対して身長と体重という二つの観測値が得られる．このようなデータを2次元データ2-dim.dataという．統計学では2次元以上のデータを1次元のデータに対して多次元データmulti-dimensional dataと呼ぶ．多次元のデータでは個々の属性の分析ばかりでなく，属性間の相互関係の分析も重要となり，分割表を使った分析や相関・回帰分析が行われる．

◆時系列データ、クロス・セクション・データ、パネル・データ

時系列データ：同一の対象の異なった時点での観測値からなるデータを時系列データ(time series data)と呼ぶ．例：1987~2000年の日本のマクロ経済データ。

クロス・セクション・データ(横断面データ)：ある時点において、複数の個体を観測することで得られるデータ．例：世界各国の2024年の人口。

パネル・データ：定めた一定範囲の対象に対して時系列データを集めたもの．例：世界各国の1987~2000年のマクロ経済データ。

◆実験データと調査データ

実験データ：実験により得られたデータ．主に自然科学の分野.

調査データ：調査により得られたデータ．主に人文・社会科学の分野.

◆全数調査と標本調査

全数調査：母集団全体を調査すること．例：国勢調査.

標本調査：標本を調査すること．例：世論調査，テレビ視聴率調査.

3　統計データの分析プロセス

- 統計データの分析はつねにデータの収集に始まる，と考えるのは誤りである．最初に行うべきことは，何を対象にどのようなことを分析するか考えるということでなければならない．分析を行うべき仮説を考えるということである．仮説がないままに，むやみにデータを集めてそれを分析しても何の意味もない.

- 仮説が構築され，分析したい対象が明らかになって，初めてデータが必要となる．もし，分析に必要なデータがもともと存在しない場合には，自らデータ獲得の作業を行わなければならない．この作業は，自然科学の分野では実験experiment，人文・社会科学の分野では調査surveyと呼ばれる.

- データが手に入れば，実際の統計分析に入る．実際の統計分析は，今日ではほとんどの場合，コンピューターを用いて行われている.

- そこで次なる段階として，人間が行う，計算された結果が統計的にどのような意味を持っているかを解釈し，それを適切に表現する手段（プレゼンテーション）を考えるというプロセスが必要となる.

- 一連の統計データ分析はこのようなプロセスを経て行われる．そして多くの場合，1 回目の分析結果を見てさらに仮説を修正するという具合に，結果をフィードバックさせながら，繰り返して分析が行われる.

4　環境構築とColabノートブック

4.1　なぜPythonを使うか

本講義では、データ分析をするためのプログラミング言語としてPythonを使います。そもそもなぜPythonを使うのでしょうか。それは、他のプログラミング言語と比べてコーディングが容易で、さまざまなこと（データの加工、取得、モデリング等）が一貫して簡単にできるからです。このようなデータ解析や機械学習系のライブラリが揃っているのが特徴です。

こうした理由で、多くのデータサイエンティストが、データ解析にPythonを利用しています。Pythonのユーザーはどんどん増えてきて、Pythonはどんどん進化しています。Pythonの構文は比較的簡単なので、Python以外でプログラムをやってきた人はもちろん、プログラム経験がない人たちでもすぐに扱うことができます。

4.2　サンプルコードを利用する際の注意点

- Jupyter Notebookという環境を使うことはできる。Jupyter Notebookのインストール方法などは配布した別紙にまとめていますので、まずはそちらを参照して環境を準備してください。なお、Googleが提供している「Colaboratory」を使ってもサンプルコードを実行できます。去年の様子を見て、、Googleが提供している「Colaboratory」を使うことをお勧めします。Googleのアカウントが必要です。

- 提示するサンプルプログラムを使って、実際に変数に入れる値を変更してみたり、コードを実行して、結果をみてください。基本的には、上から順に実行するだけで良いのですが、ただコードを眺めているだけでは、分析やコーディングのスキルは身に付きません。実際に手を動かして試行錯誤することでしか、コーディングスキルは身に付きません。

- やり方がわからなかったり、エラーメッセージなどが返ってきて、詰まる時もあるかもしれません。しかし、エラーメッセージを見ながら、まずは自分で調べながらやることも大事です。またコードが複数行あって、書籍の説明文だけでは分からない処理があるかもしれません。そのときは、1行1行実行して、どういう結果が返ってくるのか、見ていきましょう。そこから1つ1つ学ぶことができるのです。

- わからないキーワードやライブラリ名、コードなどが出てくることもあると思います。そのときは、これまであげた参考文献などを見るだけではなく、検索エンジン（Googleなど）を積極的に使って調べていきましょう。はじめは調べたいものがすぐに見つからず、時間がかかるかもしれませんが、慣れてくれば調べるコツも分かってきます。この調べる力もとても重要です。

- また、サンプルコードをすべて丸暗記しようなどとは思わないでください。あくまでも目的は、Pythonを使って、さまざまなデータ加工処理ができるということをまずは知ってもらうためにあり、すべて覚えてもらうことを想定していません。学んだばかりの処理は、すぐに使いこなせないかもしれませんが、必要なテクニックは使う頻度も多くなって、そのうち手が覚えて、自然に使えるようになります。実際、現場で働いている多くのエンジニアは、わからないことがあるときは、ネットで探したり、ネット上にある掲示板で質問して、仕事をしています。ですから特に初学者の方は、どのような方法があるというのをまず知って、必要なときに振り返って、使えることが大事です。

4.3　Colabノートブックで実装

では、Pythonのコードとは、どのようなものでしょうか。早速、Pythonのコードを見て、実行していきましょう。

まずは、プログラミング言語入門でおなじみの「Hello, world!」を表示させることをやってみます。Pythonなら、次のコードで足ります。他のプログラミング言語は数行必要ですが、Pythonではこの1行だけでよいのです。printは画面に出力する関数です。print関数の括弧中に、出力する文字列を指定します。Pythonで文字列を表現するには、「'Hello, world!'」

のように全体をシングルクォーテーション（もしくは「"」（ダブルクォーテーション））で囲みます。

Run 1: 「Hello, world!」を表示させる

```
print('Hello, world!')
```

【手順】

1. セルを追加する。新規にNotebookを作成したときは、「Untitled」という名前のファイルができ、その中に1つのセルがあるはずなので、そこにコードを記述します。もしセルがない場合、もしくは、セルを追加して他のコードをさらに実行したいときなどには、左上にある［+コード］ボタンをクリックすると、セルを追加できます。セルを追加するには、「+コード」「+テキスト」の2種類があります。

2. コードを入力する。Pythonは全角文字を正しく認識しないため、プログラムの中に全角文字やスペースが混入するとエラーが発生します。このような問題を避けるためにはプログラムの入力や編集時には、常に半角モードを使用することが重要です。#はコメントアウトで、注記をするために書いたが、コードの実行中は無視されます。今書いているコードの意味を、将来的に理解するためであったり、第3者がみてもわかりやすいように、適宜、コメントを残すことも大事です。

3. 実行する。セルをクリックして選択した状態にしておき、左端の［Run］ボタンをクリックすることでコードを実行できます。もしくは、セルを選択し、［Shift］キー＋［Enter］キーを押すことでも実行できます。［Shift］キー＋［Enter］キーで実行すると、セルが下にもうひとつ追加されて、さらにプログラムを入力できるようになります。必要なければ、ゴミ箱のアイコンをクリックして、そのセルを削除してもかまいません。

たとえば、次のコードは、足し算（+）、かけ算（*）、そして、べき乗（**）を計算します。

Run 2: 演算

```
# addition
print(1 + 1)

# multiplication
print(2 * 5)

# exponential
print(10 ** 3)
```

もっとコーディングの効率を高めたいなら、ショートカットキーを使いこなせるようになりましょう。編集モードでない状態（［Esc］キーを押します）で［Ctrl+M　H］キー押すとキーボードショットカットが示されます。以下のような画面が出てくるので、たとえば、新しいセルを下に追加したいときは、［Ctrl+M　B］キーを押します。他は、コピー(［Ctrl+C］)、貼り付け(［Ctrl+V］)などもあるので、ぜひ使いこなしてください。ショートカットキーに慣れていない人は、はじめは少し大変かもしれませんが、慣れると圧倒的に作業時間が短くなります。

今日のキーワード

統計学，記述統計学，統計的推測，推測統計学，母集団，標本，時系列データ，クロス・セクション・データ(横断面データ)，パネル・データ，実験データ，調査データ，全数調査，標本調査，1次元データ(1変量データ)，多次元データ(多変量データ)，質的変量，量的変量