統計学入門

小波秀雄

June 2022

はじめに

多数のデータから意味のある情報を抽出するのが統計的手法であり，その理論が**統計学 (statistics)** である．統計学は，確率論を基礎にして，不確実性を含む多数のデータから，一定の確実さをもった判断を下すことを目的にしている．

統計学は，社会や人間に関わるさまざまな事象の分析と多数のデータの定量的な取り扱いを可能にすることから，社会科学や医学などの人間集団を相手にした学問研究分野，心理学や教育学などの人間行動の分野，品質管理などの生産現場，保険や経営といったマネージメント分野，また政策決定のための指針作成など，さまざまの分野で広範に活用されている．

自然科学の分野でも，不確実性を含む自然現象は数多く，データの統計的な取り扱いが必要になる．また情報理論の中でも確率論とその応用は重要な一分野である．特に本書で展開される確率分布の理解は情報理論の中でも基本となるものである．

このように，統計学はまさに現代の学問と産業を支えている主要な理論のひとつであるといっても過言ではない．

その反面，確率や統計の誤った解釈や，意図的に捻じ曲げられた解釈によって，誤った指針や主張が導かれることも稀なことではない．嘘をつくための道具として，統計が不審の眼を向けられることも昔からよくあることである．誤った解釈に振り回されたり，統計の嘘にだまされたりしないためにも，統計の理論を基礎から理解することは大切である．

このテキストでは，確率論の入門からはじめて，古典統計学の理論を一通り取り扱う．数式は多いが，高校数学程度の力があれば追えるように，巻末付録に式の誘導を掲載した．あまり多くはないが，理解に必要な例題と練習問題を入れてあるので，それも含めてまじめに取り組んでいただけば，統計学の十分な基礎力を獲得できる．現代的な多変量統計や予測統計はこの先の展開になるが，その勉強のための足がかりにもなるはずである．

表計算アプリケーションの利用

統計処理では数多くのデータを使って多数回の計算を行う．

その労力を省くために，Excel や Numbers, OpenOffice/LibreOffice Calc といった表

計算アプリケーション[*1]を使うと便利だ．セルにデータを打ち込んでから，簡単な数式を使って一斉に同型の計算をさせたり，総和を取ったりできるので，このテキストの問題を解くために活用してみることをお勧めしたい．

ただし，これらを使用する際に注意しておかなければならないのは，特に統計関数を利用したときに，出てきた数字をそのまま信用してしまって，ミスを見逃してしまうことである．たとえば分散および標準偏差を求める関数として VAR と STDEV があるが，これは第 1 章で出てくる標準偏差とは定義と値が異なることを知っておかないとまずい．

統計計算のためのパッケージの利用

本格的な統計処理のためのパッケージとして，オープンソースの統計処理のためのプログラミング言語である R[*2] が開発されて，広く利用されるようになっている．これから何らかの統計処理パッケージを導入する場合には，まず R を使うことをおすすめしたい．単に「R」で検索するだけでダウンロードの仕方も含めて情報が手に入るようになっている．R については，多数の参考書やマニュアルも出版されているので，その意味でも学びやすい環境になっている．巻末に R に関する情報をまとめてあるので参考にしていただきたい．

また Python も数理分野に強く，最近は人気のあるプログラミング言語である．統計処理についても多数の参考書が出ているので，好みに合ったものを使って実務に活用することもお勧めである．

正しくアプリケーションを使うために

車を運転するのに，エンジンの仕組みや道路設計に関する知識は必要ない．それでも，どこに行こうとしてハンドルやアクセルを操作しているのかを分かっていないと，車はあらぬところに到着してしまう．ところが，それでも「目的地に到着しました！」と運転手が宣言する，そんなことがあったら客はどう思うだろうか？

ところが，「コンピュータで統計処理をやりました」といって，これとまったく同様の誤りを犯してしまうことはむしろありふれている．研究や実務に携わる人でさえ，実は統計学について無知なままに手続きだけを覚えて，結果を出しているケースは珍しくない．それを避けるには，統計的なデータ処理の意味をわかっておくことが必須であり，このテキストはそのために書かれている．

統計学を学ぶということは，難しい数学をマスターすることではないし，まして，基本的な定理の証明にまで遡って勉強する必要はないと言える．このテキストでも，ほとんど

[*1] Excel, Numbers, はそれぞれ Microsoft Office, Apple iWork, に含まれる表計算アプリケーション．
[*2] R はフリーソフト財団の GNU プロジェクトとして開発されているので GNU R と呼ぶこともある．

の数式の導出は付録に回して，数学的に納得したい人の便宜を図りながらも，本文では数学的な細部にあまり立ち入らないように留意した．

しかし，数値データを材料として処理を進める以上，その処理が何を意味しているかを理解するためには，最低限の数学的な扱いは必要である．それを押さえた上でアプリケーションの使い方をマスターすれば，安心して，かつ創造的に統計の手法を活用できる人になれるのだ．そのつもりで，本書を学んでほしい．

用語について

統計学は広範な分野で使われているために，分野ごとに，あるいは本によって用語の不統一が目立つ．このテキストでは定評のある英語の教科書とその日本語訳を中心にして，用語の統一を図った．

この本の利用について

この本の PDF ファイルは下からダウンロードできます．

`http://konamih.sakura.ne.jp/Stats/Text/`

ダウンロードは自由に行っていただいてかまいません．利用にあたっては，次の点に留意してください．

- 個人としての利用は許諾なしに行ってください．
- 学校や企業などにおける講義，セミナー等で使う際には，利用の形について著者に教えていただけると幸いです．
- 出版その他のパブリックな媒体への転載，図版の利用等については著者の許諾を得てください．
- ウェブからダウンロードできるようにするときには，古いバージョンがネット上に残ることを避けるため，上の URL へリンクすることとし，転載したファイルを別に置くことは避けてください．
- 内容に関するコメント（誤りの指摘，質問，要望など）がありましたら，メールでお知らせください．

著者連絡先

著者の肩書と連絡先は以下のとおりです．
京都女子大学　名誉教授
小波秀雄
E-mail: konami@kyoto-wu.ac.jp

9.2.4 線形回帰 (最小二乗法)

相関のあるデータを散布図にして，その中にデータの変化の傾向を表す直線を引きたいことはよくある．なお，このように全体の傾向を表す関数のことを**モデル**と呼ぶこともある[*3]．

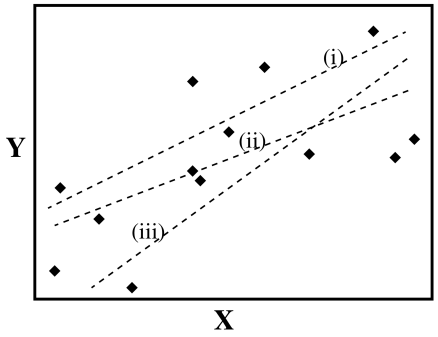

図 9.5 すべてのデータ点を代表する直線はどれがよいか

図 9.5 に描かれた 3 本の直線のうち，適切なものが (ii) であることは，勘で分かる．しかし客観的に最良のモデルとなる直線を決定するにはどうしたらよいだろうか．そのために用いられるのが**線形回帰 (linear regression)** または**最小二乗法** (**least-square method**) と呼ばれる方法である．

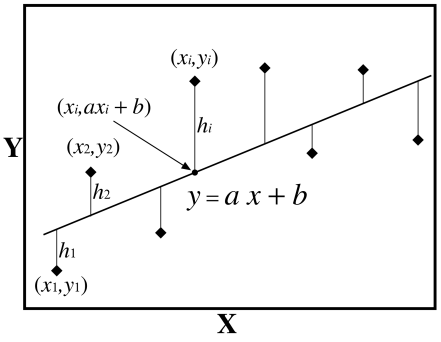

図 9.6 最小 2 乗法の原理: $h_1^2 + h_2^2 + \ldots$ が最小になるように a, b の値を決めてやる．

[*3] モデルという意味は次のようなことだ．— 与えられたデータ点の中に潜む関係としてはいろいろなものが考えられる．その中で単純な線形な関係を，仮にひとつの「モデル」として仮定してデータをそれに当てはめることで，問題を分析したり解釈したりしたい．

図 9.6 のようにデータ点 $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ が与えられていたとき，$y = ax + b$ で表される直線を引いたとしよう．このとき，i 番目のデータ点と直線の縦のずれを h_i とすると，

$$h_i = y_i - (ax_i + b) \tag{9.10}$$

となる．直線 $y = ax + b$ は，a と b の値を変化させることで，傾きを変えたり平行にずらしたりできる．それでは a, b がどのような値をとったときに，直線はデータ点をもっともよく近似できるだろうか．詳しい導き方は付録（p.167）にゆずり，ここでは結果だけを示す．

$$a = \frac{\overline{xy} - \overline{x}\,\overline{y}}{\overline{x^2} - \overline{x}^2} = \frac{\sigma_{xy}}{\sigma_x^2} \tag{9.11}$$

$$b = \overline{y} - a\overline{x} \tag{9.12}$$

式 (9.11),(9.12) で得られた a, b を用いて直線を引くと，データの増減を「ほどよく」表した直線が得られる．このような直線を**回帰直線**と呼ぶ．

例題 9–2　成人男子 6 人の靴のサイズと身長を調べたところ，次のようなデータの組が得られた．これらを散布図にプロットし，最小 2 乗法を使って回帰直線の係数を求めて，図に直線を描きなさい．

$(24.5, 165.4), (28.0, 182.7), (26.0, 171.6), (25.5, 173.1), (25.0, 175.1),$
$(24.0, 170.6)$

これらのデータの組を $(x_1, y_1), (x_2, y_2), \ldots$ とすると，式 (1.6) から分散 σ_x^2 が，式 (9.2) から共分散 σ_{xy} が得られる．具体的には x_i，y_i，x_i^2，y_i^2，$x_i y_i$ それぞれの平均，$\bar{x}, \bar{y}, \bar{x^2}, \bar{y^2}, \bar{xy}$ を個別に計算してから分散と共分散を求め，式 (9.12) に代入すればよい．電卓を使ってもかなりめんどうなので，Excel などを利用するとよい．

データの散布図と，このようにして求めた a, b を使って引いた直線を図 9.7 に示した．

以上のようにして求められた回帰直線で示される相関が，真の相関であるのか，あるいは母集団には相関がないのに，抽出によってたまたまある傾向が現れてしまったのかという疑問は，特にデータの点が少ないときには問題になる．このことについては次節で考えてみよう．

回帰直線\Statistics 小波秀雄 209p.pdf

付録 A

重要な関係式などの導出

A.8　最小二乗法

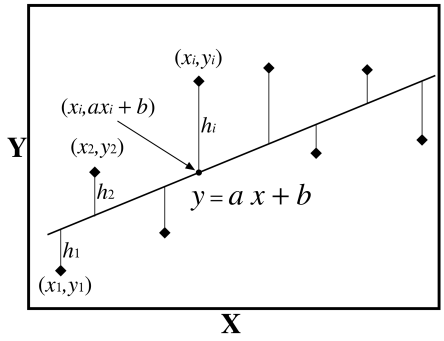

図 A.2　最小二乗法の原理: $h_1^2 + h_2^2 + \ldots$ が最小になるように a, b の値を決めてやる.

図 A.2 のようにデータ点 $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ が与えられていたとき，$y = ax + b$ で表される直線を引いたとしよう．このとき，i 番目のデータ点と直線の縦のずれ

を h_i とすると，

$$h_i = y_i - (ax_i + b) \tag{A.36}$$

となる．直線 $y = ax + b$ は，a と b の値を変化させることで，傾きを変えたり平行にずらしたりできる．それでは a, b がどのような値をとったときに，直線はデータ点をもっともよく近似できるだろうか．

そのためにまず次の S を定義しておこう．近似がもっともよいときには，S は極小になるはずである．

$$S = \frac{1}{n}(h_1^2 + h_2^2 + \ldots + h_n^2) \tag{A.37}$$

ここで右辺に $\frac{1}{n}$ を掛けているのは，後の計算をうまく処理するためである．

式 (A.37) に式 (A.36) を代入して整理すると，次の式が得られる．

$$S = \overline{y^2} + a^2\overline{x^2} - 2b\overline{y} - 2a\overline{xy} + 2ab\overline{x} + b^2 \tag{A.38}$$

a, b を変化させて S が最小になる条件を求めるには，次の 2 つの偏微分がゼロになればよい．

$$\begin{aligned} \frac{\partial S}{\partial a} &= 2a\overline{x^2} - 2\overline{xy} + 2b\overline{x} = 0 \\ \frac{\partial S}{\partial b} &= -2\overline{y} + 2a\overline{x} + 2b = 0 \end{aligned} \tag{A.39}$$

これを整理して，次のような連立方程式が得られる．ただしここでは a, b が未知数であることに注意！

$$\overline{x^2}a + \overline{x}b = \overline{xy} \tag{A.40}$$

$$\overline{x}a + b = \overline{y} \tag{A.41}$$

これを解くと次の結果が得られる．

$$a = \frac{\overline{xy} - \overline{x}\,\overline{y}}{\overline{x^2} - \overline{x}^2} = \frac{\sigma_{xy}}{\sigma_x^2} \tag{A.42}$$

$$b = \overline{y} - a\overline{x} \tag{A.43}$$