

目次

第 1 章	縦割りの統計学	2
1.1	記述 vs 推測	2
1.1.1	二項対立の構造	2
1.1.2	回帰の例	2
1.1.3	仮説検定	2
第 2 章	統計的推論	3
2.1	頻度派	3
第 3 章	多変量解析	4
3.1	歴史	4
3.2	多変量解析の考え方	4
3.3	用語と記法	4
第 4 章	時系列解析	5
4.1	歴史	5
4.1.1	パラメータ付け	5
4.1.2	推定法	5
4.2	基本的性質	5
4.2.1	エルゴード性	6
4.3	定常過程	6
4.4	スペクトル解析	6
4.5	リード・ラグ効果と CCK 理論	6
参考文献		7

第 1 章

縦割りの統計学

1.1 記述 vs 推測

1.1.1 二項対立の構造

記述統計学と推測統計学との手法の違いは、データの背後に確率論的な構造を考えるかどうかという点にある。データを所与のものと思い、データに対する理解を深める手法と、データを確率変数の実現値と捉えて予測を目指す手法とである。

これはアルゴリズム vs 推論という二項対立でもある。汎関数の計算・実装と、推定である。現在は前者の黄金期であり、推論法に進化を要求する。アルゴリズムは統計学の外に触発される、神経回路網、サポートベクトルマシン、ブースティングなど。そしてこれを受け止めるために数理がさらに豊かになる。

1.1.2 回帰の例

線型回帰の最小二乗法は古典的なアルゴリズムである。その推定の正確性を評価する推論法には、標準誤差などがある（95% に入る振れ幅）。近年のアルゴリズムには lowess (locally weighted scatterplot smoother) がある。これには ± 2 倍のブートストラップ標準誤差などで推論される。ブートストラップ法は解析的な方法ではない（標準誤差のように式を持たない）が、それゆえにいかなる複雑なアルゴリズムにも適用できることが有用になる、計算量にものを言わせる脳筋手法である。

1.1.3 仮説検定

推定量の評価以外にもう一つ、推論法が要請される分野は、仮説検定である。2 標本 t -統計量 $t = \frac{\bar{x} - \bar{y}}{\widehat{\text{sd}}}$ は、ステューデントの t 分布が帰無分布となる。

第 2 章

統計的推論

観測データの背後に確率モデルを想定する仮定から始まる．

2.1 頻度派

20C から統計学の営みが開始され，卓上計算機でも実行可能なアルゴリズムによる理論が完成した．これを古典理論，または，頻度派的な推論という．

第 3 章

多変量解析

3.1 歴史

- (1) 現代的な数理統計学の理論体系が確立したのは 1920s の Fisher, Neyman, E. S. Pearson らによる。
- (2) この枠組み（相関分析など）に沿って，30s からは自然な形で多変量に拡張された．Fisher 自身も多くの貢献をしたが，Hotelling, Wilks, Wishart らの仕事が基礎になっている。
- (3) 多変量推測統計の Anderson による標準的教科書 "An Introduction to Multivariate Statistical Analysis" は 1958 年に刊行された．この頃には多変量正規分布と線型モデルに基づく推測理論の基礎は確立していた。
- (4) 1970s 以降の計算機時代では，今日ではほとんどのデータは多変量データであり，取り立てて多変量分析と言わなくても，新たな統計手法は押し並べて多変量データを扱う手法となっている．今日の統計パッケージの中にある探索的データ解析や射影追跡などは，正規分布を前提としない，非線形的手法である。

[1] の序文．

例 3.1.1. 回帰分析，分散分析，主成分分析，判別分析，因子分析，分割表，グラフィカルモデル．

3.2 多変量解析の考え方

多変量解析においては，連続分布としては多変量正規分布以外に扱いやすい分布が少ない．ところが，実際のデータ解析の場面では多変量正規分布の仮定の妥当性が疑われる場面が多く，記述統計的な手法と推測統計的な手法の間に大きなギャップがあるので現状．[1]

3.3 用語と記法

行列やベクトルを言葉として用いるのが特徴であり，経済学はこの Anderson らの慣習を強く受け継いでいる．

定義 3.3.1 (data matrix, sample size, dimension).

- (1) 多変量データをデータ行列とも呼ぶ．一般には行を個体，列を変数とする．それぞれの添字を t と i, j と使い分ける．
- (2) $n \times m$ データ行列の行数 n を標本の大きさといい，列数 m を標本の次元という．^{†1}
- (3) 値が数ではない変数を質的変数という．これを量的変数に変換する手法を数量化という．

注 3.3.2 (spatial statistics, random field). データ行列の古典的な枠組みに入らない多変量データには，地理的なデータなどがある．これには空間統計や確率場の手法がある．また，質問表に分岐がある場合は，特定の回答をした者以外の値が欠測値となる．

^{†1} 高次元データを，なんらかの形で 2 次元空間に写して解釈する記述統計的手法を，次元の縮約という．

第 4 章

時系列解析

確率変数列 X_1, \dots, X_n の実現値と見做し得るデータに対する統計解析を多変量解析と呼ぶならば，確率過程の実現値と見做し得るデータに対する統計解析を時系列解析という．そこで，

4.1 歴史

- (1) 1940s より Wiener や Kolmogorov により弱定常過程のスペクトル解析と予測の理論が確立される．特に連続時間の確率過程の理論の精緻な発展の原動力となった．
- (2) 1976 年刊行の George Box と Jenkins による "Time Series Analysis, Forecasting and Control" が，自己回帰移動平均 (ARMA) モデルと呼ばれる線型モデルの構築と予測のための標準的な手続きが確立した．これ以降は ARMA モデルの限界を乗り越えることを念頭に，非線形モデルなどさまざまな手法が提案されることになる．

通常の統計学は主に独立標本に対する議論であるが，時系列解析は時間軸の間にも従属関係がある状況での統計解析であり，その点で一般化であると捉えられる．

4.1.1 パラメータ付け

まず，時系列データを定量的に特徴付ける (パラメータ付ける) 手法を考える．

- (1) 時間領域法：自己相関など．
- (2) 周波数領域法：スペクトル解析など，

定義 4.1.1 (autocorrelation function). 確率過程の共分散関数の概念を自己相関または自己共分散といい，相関係数版の概念を自己相関関数という．

注 4.1.2. しかし，実際のデータから推定しにくい，視覚化しにくいなどの難点がある．

4.1.2 推定法

等間隔標本 $(X_t)_{t \in \mathbb{Z}}$ の i.i.d. を仮定すると Glivenko-Cantelli の補題から一致推定が可能であるが，一般には他の仮定を必要とする．

4.2 基本的性質

i.i.d. 以外に等間隔標本 $(X_t)_{t \in \mathbb{Z}}$ にどんな仮定をおけるかを考える．

4.2.1 エルゴード性

記法 4.2.1. はじめの n 回の観測からの標本平均を $\tilde{x} = \tilde{x}_n := n^{-1} \sum_{t=1}^n X_t$ とする.

定義 4.2.2 (EPCL: ergodic property with a constant limit).

$$\exists_{\mu \in \mathbb{R}} P \left[\lim_{n \rightarrow \infty} \tilde{x} = \mu \right] = 1.$$

4.3 定常過程

4.4 スペクトル解析

4.5 リード・ラグ効果と CCK 理論

定義 4.5.1.

- (1) 2つの時系列がタイムラグを持って相関する現象を lead-lag 効果という.
- (2)

要諦 4.5.2. 現状の取引の流動性では, ms の世界で存在する. 結果, 80s までは分のスケールで観測されていた現象だが, 現在は高頻度データと呼ばれる分野となってしまった. そこで, 古いモデルが使えなくなってしまったこの現象に対して, Hoffmann, Rosenbaum, and Yoshida (2013) が新しいモデルと推定法を提案した.

参考文献

- [1] 『統計科学のフロンティア 1 統計学の基礎 I 線型モデルからの出発』(第 I 部「多変量解析入門」(竹村彰道), 第 II 部「時系列解析入門」(谷口正信))