統計的諸手法に現れるロバスト性の概念

数理情報学専攻 竹村彰通

2006年1月

概 要

ここでは，統計学の考え方について歴史的な展望を含めて総括するとともに，統計的手法に現れるロバスト性のさまざまな側面について説明する．

1 統計学の流れ

統計学は様々な起源を持つ学問である．そもそも，統計学とよばれる共通な方法論が独自の分野として認識され分野として確立したのは20世紀の前半になってからであり，それまでは統計的方法はそれぞれの応用分野で個別に発達して来た．

統計学の起源としては，1) 社会経済的な起源，と 2) 科学技術的な起源，を分けることができる．まずは前者について述べる．

統計学 (statistics) の語源は「国 (state)」にあり，税金や軍事などの目的のために国力を定量的に評価することを中心とするものであった．また大航海時代からはじまった海難事故等にそなえる損害保険の発展は，統計的なリスク共有という概念をもたらした．さらにはグラントの死亡表 (1662) にはじまる生命表の発展は生命保険につながっていった．国力の定量的な評価に関してはペティの『政治算術』(1690) が歴史的に重要である．またケトレー (1796–1874) は社会の大量観察を重要視し,「平均人」の概念を提唱した．以上のような統計学の社会経済的な起源は，数量的な側面を重視するものであり，その意味では次に述べる科学技術的な起源とも共通の側面を持つものであった．

科学技術的な起源としてまずあげられるのは，フェルマーとパスカルの往復書簡 (1652) によってテーブルゲームと賭けの文脈で確率論が論じられ始めたことである．その後，ベルヌーイ，ラプラス等によって大数法則，中心極限定理などの基本的な確率論の定理が整備されていった．また天体観測にともなう誤差解析については，ルジャンドルやガウスによる最小二乗法や正規分布の発見が重要である．さらに19世紀末から20世紀初頭においては，遺伝法則の発見とそれにともなう生物測定の分野で，ゴルトン，カール・ピアソン等によって相関や回帰分析などの統計的手法が発展した．

以上のような背景の中で20世紀になると，主にイギリスにおいて，汎用的な方法論としての統計学の独自性が主張されるようになった．カール・ピアソンは「科学の文法」と

しての統計学を提唱し，また R.A. フィッシャー (1890-1962) は推測統計学，実験計画法，最尤法などの統計学の重要な手法を確立した．実験計画法は，品種改良などの農業実験の場から発達したことは興味深い．

統計学の分野としての成立に並んで，コルモゴロフ (1933) による測度論的確率論の確立の重要性についてもふれておく．測度論的確率論においては，確率を公理的に構成し，確率の意味を問わないことが特徴となっている．そして極限操作等の数学的操作の正当性が保証された．これにより多くの分野での確率論の応用の基礎が固まったと言える．

戦後になると，統計学の中心はアメリカに移り，数学の一分野としての「数理統計学」の形式が整えられることとなった．これは当時の数学の抽象化の流れにそうものであるが，一時やや抽象化が行きすぎたきらいがある．一方で統計的手法の応用分野はさまざな拡がりを見せた．

「文系分野」としてはまず計量経済学の発展があげられる．これは経済を連立方程式システムとしてモデル化して，システムの統計的推定をおこなおうとするものである．またアメリカにおける数量的な研究重視の背景の中で，心理学，社会学などでもさまざま統計的手法が開発されてきた．

「理系分野」としては，医学統計 (大規模臨床試験，新薬の認可等) の発展が現在でも重要である．また実験計画法の工業への応用では，統計的品質管理がわが国の工業製品の品質向上に重要な役割を果たした．田口玄一による「田口メソッド」においては，品質のばらつきに影響を与える要因の分析が強調され，さまざまな使用環境で安定的な品質を保つ製品の開発のために,「ロバスト設計」の考え方が提唱され大きな影響を与えた．より最近では，金融工学や数理ファイナンスにおいて，伊藤積分などの抽象的な確率解析の手法が重要な役割を果たした．

2　統計的手法の変化

前節て述べたような統計的手法の応用分野のひろがりとともに，統計的手法の性格も大きな変化をとげた．その背景として，計算機の発達により，実際に複雑な計算ができるようになったという事実がある．そして，用いられる統計モデルが大規模化していった．

そのなかで，モデル選択の方法論が重要となった．まず最尤法を基本とするモデル選択について述べる．統計的モデル選択においては，モデルの説明力 (データへのフィット) とモデルの簡潔さのバランスをとることが重要である．つまり，複雑なモデルは手もとのデータへのフィットは高いが，新たなデータの予測が不安定 (overfitting, 「過学習」) となることが多い．モデルの簡潔さはロバスト性と言ってもよい．この問題に対して，AIC (赤池情報量規準) などの手法によるモデル選択が有用である．この分野は我が国の貢献が大きい．

最尤法をベースとする統計的手法と並んで，最近ではベイズ法が重要となっている．ロバスト性の観点から見ると，ベイズ法はロバストな側面とそうでない側面の双方を含んで

いる．ベイズ法では，システムのパラメータに関する不確実性や，分析者の主観的な情報などをすべて確率変数としてモデル化する．これを事前分布とよぶ．システムの変数の値を決め打ちせず，確率変数としてその不確実性を考慮するという意味では，ベイズ法はロバストな側面を有する．しかしながらベイズ法では，事前分布を特定化してその事前分布を前提とした最適化をおこなうことから，事前分布の特定化の段階においてロバスト性が失われる可能性がある．

さて，ベイズ法ではシステムの未知変数は単に観測されていない確率変数であると考える．さらに，統計的推論は，データに基づいて事前分布を事後分布でおきかえる機械的な操作となる．これにより，統計的推測はすべて事後分布を求める確率計算に帰着することが特徴的である．このため，機械的な操作で統一的な観点から統計的推論をおこなえるという点で，コンピュータとの相性がよい．また，観測されない「潜在変数」の扱いも比較的容易で，扱えるモデルのクラスが広い，という利点も有している．このためにベイズ法は一種のパラダイムとして様々な分野で利用されるようになってきている．

さらに，ベイズ法における事後分布計算のためのモンテカルロ手法の発展も重要な要因である．モンテカルロ法は，確率計算にともなう積分などの計算を，乱数を用いて近似的におこなうものである．近似の誤差があり，また計算の効率性は必ずしも高くないが，汎用性があり実装が容易である点が利点である．

特にマルコフ連鎖モンテカルロ法 (MCMC) とよばれる手法が発展しており，これによりモンテカルロ法の適用範囲は大幅に広がった．MCMC 法は，ある確率モデルに従う乱数を発生させたいものの，モデルに従う乱数を正確かつ直接に発生しにくいような状況で有効である．単純化して言うと，MCMC 法は，ある確率モデルに従う乱数を近似的に発生させる手法であり，モデルを解く際の収束計算のステップを乱数をもとにおこなうことにより，目的とする確率モデルに従う乱数を近似的に発生する．つまり，乱数発生と収束計算の手法を組み合わせることにより，モンテカルロ法の適用範囲を大幅に広げるものである．

以上で，最尤法およびベイズ法について述べたが，これらとはまた別の考え方として，統計的モデルを有限次元のモデルに特定化しないノンパラメトリックなアプローチも重要である．このようなアプローチとして，推定方程式法あるいは一般化モーメント法とよばれる方法がさかんに研究されている．独立成分分析とよばれる手法も推定方程式法の一つとみなすことができる．推定方程式法は，確率分布に対して直接の仮定をおかず，変数間の制約条件にのみパラメータを入れ，制約条件のみからパラメータを推定する方法である．このため，ロバスト性の観点からは非常にすぐれた推定法ということができる．推定方程式法は，計量経済学において多用されている．

3　統計的方法の性格

ここでは，以上のような概観にもとづき，統計的手法の性格について述べる．統計学では，一定の均一性を持つ母集団を想定する．ただし，母集団におけるバラツキの存在を前

提とする．例えば生命保険の例で考えよう．生命保険では，被保険者を性別，年齢などの一定の条件で分ければ，均一な集団となると考え，その集団に一定の保険料を設定しリスクをプールする．このような集団の均一性を信じるには情報が少ないほうがいいこともある．例えば，個人の遺伝子情報などがとれて情報が多くなりすぎると，特定の個人は特に高いリスクを持つことがわかってしまうなど，集団の中でリスクをプールすることができなくなる可能性がある．このような場合，通常の保険の考え方では対処できず，社会保障のような別の考え方をとる必要が出てくる．

また，異なる例として，じゃんけんで人間に必ず勝てる機械が話題となっている．じゃんけんは手軽なランダム化のメカニズムとして機能しているが，人間の手の動作を非常に早く判読できる機械が出てくると，その機械は人間に勝ててしまう．つまりじゃんけんのランダム化の機能が失われる．これも過剰な情報が利用できるために，ランダムネスが失われる例である．

いずれにせよ，統計的手法を応用するには，一定の重要な要因を所与とすると，それ以上は純粋な確率的な変動であり，説明できないとわりきって考えられる点が重要である．これを 「固定効果」+「誤差」 という形で表現することがある．また，確率的な変動は多数回反復可能であり,「リスクをプールする」ことにより，平均的な性能が支配的となる状況を設定する．

ここで，薬の例をとって「平均的な性能が支配的となる状況」の意味を説明しよう．薬として，(その効果は劇的でなくても) 平均的に効く薬がいいのか，それとも特定の状況で劇的な効果をもたらす「特効薬」がいいのか，という問を考える．統計学のとりあえずの答は，どのような条件の人にも平均的に効く薬がいい，というものである．これはロバスト性のある薬がよい，と言ってもよい．あるいは統計的な用語を用いると，交互作用のない薬がいいと言うこともできる．さらに，特定の状況で効く特効薬といっても，その状況がある程度の頻度で起きるものでなければ意味がないとも言える．その意味では，あくまで平均的な性能が重要である．

以上から，統計的方法の性格をまとめると，

1) 統計モデル = 固定効果 + 誤差 の図式が成り立つと仮定し
2) 誤差はブラックボックスと仮定し,
3) 誤差の分布について平均的な性能で評価し,
4) 固定効果の不確実性は未知母数によって処理する

と述べることができる．

以上のような考え方に基づく統計的手法は，最近では様々な場面で汎用的な方法として有用性が認識されている．この考え方を極端に押し進めると，すべてを確率計算に帰着させようとする方法論となる．すなわち，定数も分散の小さい確率変数であり，システムの未知母数も確率変数であると考えれば，ベイズ法によってすべては確率計算に帰着するからである．データが大量にとれる場合には，このような統計モデルによるアプローチは,

多くの場合他の方法より有用である．その理由として，統計的方法が本質的にブラックボックス的アプローチであり，ブラックボックスの意味を問わないという実用主義にたっているという点があげられる．

ただし，現象のどの部分をブラックボックスとし，どの部分について「意味を考えるか」という視点は重要である．現象の意味を考えるには，それぞれの個別分野の知見を必要とする．統計的手法はそのような知見の及ばない部分を確率的に扱う手法と考えるのが，伝統的でありやはり健全であると思われる．