統計学とは

1 統計学 (statistics) とは？

- ガリレイ (Galileo Galilei; 1564-1642): 「自然は数学という言語で書かれた書物である。」
- カール・ピアソン (Karl Pearson; 1857-1936): 統計学は「科学の文法である。」
- **統計学** (statistics) は過去 2 世紀以上もかかって、多くの分野と係わり合って、できた学問である。
  1. ゲームのテープルから起った 確率論;
  2. 国家財政上の必要から起った 国家状態の統計;
  3. 難波事故や海上掠奪に対する 海上保険 の計算;
  4. 17 世紀のペスト禍を機とする近代 死亡率表 の研究;
  5. 天文観測で生じる 観測誤差の理論;
  6. 生物等で生じる諸量の 相関関係の理論;
  7. 農学で実験を計画するための理論として知られる 実験計画 の理論;
  8. 経済学や気象学における 時系列 の理論;
  9. 心理学における 要因分析 や ランキングの理論;
  10. 社会学における χ^2 統計量の方法;
- **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。
- **記述統計学** (descriptive statistics) 現象の法則性を知るために、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を見出す理論を**記述統計学**という。
- **推測統計学** (inferential statistics) **確率論**という数学の理論を武器として、記述統計学の上にここ一世紀ほどで打ち建てられた方法論の体系が、**推測統計学**である。

  近代統計学は、記述統計学と推測統計学をあわせたものといえる。

1.1 近代統計学の成立

生物測定学 biometry が近代統計学理論の発展の全面を担った（になった）。ここで、**相関**と**回帰**という統計学上の重要な方法を例にとって説明する。ゴルトンは有名な遺伝学者であった。彼によるスイートピーの種子の直径の測定では、親を横軸に子を縦軸にとると、データは大体傾き 1/3 の直線の近くで分布しており*、親がばらつくほど子はばらつかない。全体として、みな平均に退行（**回帰**）してゆく。**相関**や**回帰**が、法則性の表現として歴史的に初めて意識的に用いられた例である。

このように歴史的に、回帰という言葉は、

*最小自乗法による直線の傾き推定量は 0.343 である。

遺伝を繰り返すと子の特徴が平均値に近づく

という現象を説明するのに用いられた。この現象を表す手段として直線回帰が使われたことから、回帰という言葉が使われるようになった。つまり、本来の回帰は、直線への回帰ではなく、平均値への回帰、すなわち、将来の親が生む子がより平均値に近い特徴を持つことを意味する。

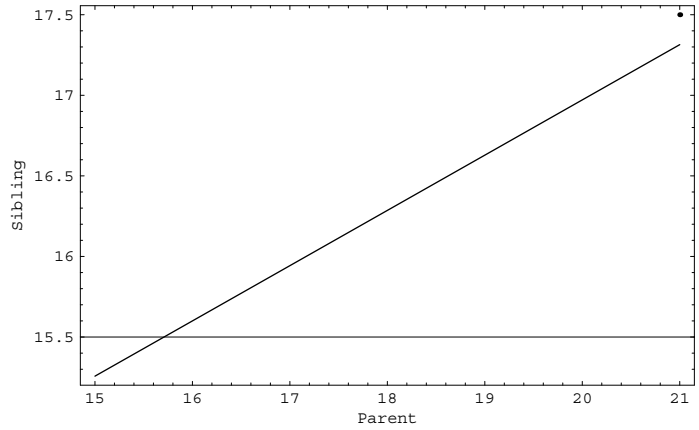

図 1: スイートピーの種子の直径に見られる、平均への**回帰**の傾向。直線は最小自乗法によって得られている。

相関係数の概念は、後に K. ピアソンによって次のように厳密に定義されている

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{1.1}$$

ただし、$\bar{x} = n^{-1}\sum_{i=1}^{n} x_i$, $\bar{y} = n^{-1}\sum_{i=1}^{n} y_i$ は標本平均を表している。式 (1.1) はしばしばピアソンの**積率相関係数** (product moment correlation coefficient)、或いは単に、**相関係数** (correlation coefficient) と呼ばれている。

2 統計学の発展の歴史

2.1 中世までの統計学

最古の統計資料としては，家畜数や財産を記録するために木に刻みつけられた原始人の印等があるが，史実に残っているものとしては，紀元前 31 世紀頃古代エジプトでピラミッド建設のために行われた統計調査がある．またエジプトでは，紀元前 22 世紀・14 世紀頃にも土地調査が行われている．一方中国では，紀元前 20 世紀頃の殷王朝時代に行われた国勢調査がある．また旧約聖書には，紀元前 15 世紀頃イスラエルで行われた国勢調査が引用されている．インドでは紀元前 4 世紀以前から，税金の納付状況，各カーストの住民数，職種別労働者数，家畜数などの行政記録がとられ，確保できる労働力や課金について把握したとの記述が見られる．古代社会における人口調査・土地調査はこれら以外にもたくさんの記録があり，例えばペルシャの国勢調査，スパルタやアテネの租税表・財産簿作成のための国勢調査等がある．

日本では紀元前７世紀や紀元１世紀に人口統計調査が行われたと言われているが，史実に残る古い記録としては，７世紀の班田収授法や庚午年籍に見られる人口や土地に関する調査がある．

census(国勢調査) という語は，ラテン語の censere(税金) から派生したもので，もともとは財産評価を意味していた．定期的に国勢調査が行われるようになったのはローマ帝国で，紀元前５世紀になってからであり，課税や兵士数の把握のために，人や財産の登記簿を５年ごとに作成した．しかしローマ帝国の没落後には，ヨーロッパにおける国勢調査の記録は少なく，9 世紀頃のドイツにおける王領一覧表の作成，11 世紀のイギリスにおける王国土地台帳の作成等，二・三のものを除いてほとんど見あたらなくなる．インドでは，収穫物の収量や価格，職業別人数や賃金，種々の食品・衣類・アクセサリー・家畜などの平均価格等々に関する大規模な統計調査の記録が残っている．

以上のように，この時代までの”統計”は，基本的にはすべての人やものを数え上げることだけであり，行政者が国力の現状を把握して課税や徴兵のために役立てることが目的であった．

2.2 近代における統計

16 世紀に入って，イタリアやフランス，オランダでは，国家状況の系統的・体系的記述を目的とした国状学が発展してきた．17 世紀半ばにはドイツの大学教授 (ドイツ大学統計学派) により，国状学 (国勢学) は学問的に整備された．それは大量観察に基づく数量的記述ではなく，国家の安寧に関わる顕著事項だけを記述するというものであり，今日の官庁統計の内容に近いものであった．現在統計学は ”Statistics” とよばれているが，これはアッヘンワール (Achenwall,G.) が国状学の学問名として ”Statistik” を用いたのが始まりとされている．

一方イギリスでは，哲学者ベーコン (Bacon,F.) の影響を受け，社会現象を大量観察して数量的資料に基づいて法則性を発見しようとする政治算術学派が誕生した．このような潮流の中で，ロンドンの商人グラント (Graunt,J.) が出現し，それまでの”統計”の考え方に転機をもたらした．グラントは 1662 年に「死亡表に関する自然的および政治的諸観察」を著したが，それは市販の莫大な量の死亡表を精密に観察して数枚の表に要約し，それから導ける人の出生・死亡に関する法則を発見したものである．このようにグラントは，”統計”が単なる数え上げではなく，大量のデータを要約して有用な情報を抽出し，それから自然的・社会的法則を発見し，将来の指針を決定できるという”統計学”の考え方の有用性を実証して見せた．グラントと親交のあった財政経済学者のペティ (Petty,W.) は，ロンドンの人口予測やヨーロッパ諸都市の死亡表の比較検討を行っている．死後には，グラントの方法を適用して政治・財政の諸問題を実証主義的に検討した「政治算術」が刊行されている．また人口統計の分野でも，グラントやペティの方法論を引き継ぎ，ハレー (Halley, E.) による「人類の死亡率推算」(1693 年)，ジュースミルヒ (Süssmilch,J.P.) による「神の秩序」(1741 年) 等が刊行された．この頃までには，比率・平均・中位数などの統計用語や，大数の法則につながる考え方が既に誕生していた．

19 世紀に入ると統計学の数学的定式化が進展し，ベルギーの天文学者・数学者であるケトレー (Quetelet,L.A.J.) は 1835 年に「人間について」を刊行でし，近代統計学への次のステップを築いた．彼は，人間に関する現象の中に法則を発見するためには多くの数を観察して帰納的に推論すべきことを主張し，平均の重要性を唱え，犯罪数や犯罪割合に関する社会的法則などを発見した．これとほぼ同時代に，看護婦の社会的地位を確立したナイチンゲール (Nightingale, F.) は，病院の統計的データ分析を行い，入院患者の死亡率を減少させることに成功している．

この頃から，各国において定期的な国勢調査が行われるようになった．イギリスでは 1801 年に，フ

ランスでは 1876 年に，またロシアでは 1896 年に第１回の国勢調査が行われている．また各国の統計局開設は，アメリカが 1790 年，フランスが 1800 年，ベルギーが 1831 年であったが，日本は鎖国の影響で 1871 年とかなり遅くなっている．第１回の国際統計会議が開催されたのは 1854 年であり，現在の国際統計協会 (ISI) が設立されたのは 1885 年である．

ケトレー以降の近代統計学の流れの中では，まず経済学と計量生物学が統計学の主要な研究対象分野であり，経済学ではイギリスのボーレイ (Bowley,A.L.) やユール (Yule,G.U.)，ドイツのレキシス (Lexis,W.) 等が，また計量生物学ではゴールトン (Galton,F.) やカール・ピアソン (Pearson,K.) 等が先駆者達であった．特にゴールトンは相関や回帰の概念を導入し,「遺伝的天才」(1869) や「指紋」(1882) 等を著し，優生学を創始した．現在記述統計学とよばれている分野はゴールトンに負うところが多い．記述統計学と推測統計学を最初に結びつけようとしたのは，ゴールトンの弟子であった K. ピアソンであり，モーメント法やカイ２乗検定などを発見した功績は極めて大きい．彼はまた，生物学者ウェルドン (Weldon,W.F.R.) と共同で雑誌”Biometrika”を創刊した．

2.3　20 世紀以降の統計学

記述統計学は今日でもなお重要なものであるが，その最盛期は K. ピアソンの時代であった．ゴセット (Gosset,W.S. ; ペンネームはスチューデント [Studet]) は，1908 年に t 分布を発見したが，この論文はフィッシャー (Fisher,R.A.) を強く刺激した．フィッシャーは K. ピアソンの理論を改良し，最尤法によるパラメータ推定と未知パラメータを推定した場合のカイ２乗検定の自由度についての研究を行ったが，特に 1922 年の推定論の論文は，その後の理論統計学の基礎を築いた．フィッシャーの影響により，有意性検定の重要性も認識され，ホテリング (Hotelling,H.)，ボース (Bose,R.C.)，ウィルクス (Wilks,S.S.) 等により標本分布論に関する多くの研究成果が得られた．フィッシャーはまた，実験計画法の発展にも貢献した．その後確率論の進歩とも相まって，ネイマン (Neyman,J.) や E.S. ピアソン (Pearson, E.S.) により，特に仮説検定論の理論体系が構築され，多くの優れた研究成果が得られた．これらの研究はワルド (Wald,A.) に引き継がれ，統計的決定理論へと発展した．

1940 年代と 50 年代には確率分布とその特性についての詳細な研究が行われ，1960 年代にはロバスト統計が研究され始めたが，これは感度解析やセミパラメトリック理論と密接な関連を持って進展している．1970 年代になると一般化線型モデルが登場し，ガウス以来の正規性の仮定を弱めることに成功した．これはまた，2 値データの解析や確率過程論とも相まって，医学データの解析において重要な生存時間解析法へつながっていった．70 年代にはもう一つの重要な展開であるデータ解析の出現があった．データ解析には記述統計に基づくものと推測統計に基づくものとがあるが，多くの場合推測的データ解析が行われている．これに対してテューキー (Tukey,J.) は，探索的データ解析 (Exploratory Data Analysis ; EDA) を提案した．テューキーは，データは様々な角度から眺めてその特徴・特異性を見いだすべきであり，そのためにはデータを要約するロバストな統計量や計算機を用いた図的表現を活用すべきであると主張した．1980 年代には漸近理論に関する様々な研究が活発に行われる一方，コンピュータを活用する統計的方法論が登場した．時代の要請もあり，ブートストラップ法，射影追跡法，回帰関数の推定法等の，ノンパラメトリックな方法論が登場した．1990 年代になってからは，ニューラルネットワークや状態空間モデル等の研究が盛んに行われるようになっており，90 年代に数理科学の一分野において起こった「複雑性」の研究にも影響を及ぼしている．さらにマルコフチェイン・モンテカルロ (MCMC) 法などの再発見により，ベイズ統計も再興しつつある．

2.4 今後に対する展望

安価で高性能なコンピュータの出現は，多くの科学や技術に大きな影響を与えており，統計学もその例外ではない．少し前までは対象とする統計モデルは，現実の現象をある程度まで忠実に記述してはいるが，数学的に解析可能なものに限定せざるをえなかった．しかしコンピュータの登場により，できるだけ現実に忠実な精密なモデルを構築し，シミュレーションにより解析を行うことが可能になってきた．このような状況の下で出現してきた統計解析の手法には，例えばブートストラップ法，MCMC 法，ノンパラメトリック平滑化法などがある．またコンピュータの発展とともに，EDA 的な考え方も拡がっていくと考えられる．さらにグラフィカルモデルの展開も，統計学の新しいパラダイムとして注目される．

上述したように，1970 年代以降データを研究対象とする分野が出現したが，それは統計科学以外にも，電子工学，制御システム工学，情報理論，人工知能論，ニューラルネットワーク，情報処理等広範囲に及んでいる．これらの分野におけるデータに対する接近法は，統計学とは異なる側面も有しているが，コンピュータ科学等の台頭と相まって，統計科学を変化させつつあることは事実であろう．統計的推測は基本となる確率モデルの設定から出発するが，これらのアプローチの中にはランダムネスを全く考慮せずに行われるものもある．しかしこれによってデータの解析の自由度は増しているとも考えられる．最近，”データマイニング” とよばれる，統計学を背景に持たない大規模データを解析するための方法論が提案されているが，将来統計学とどのように関連しあうのかに注目する必要があろう．一方ニューラルネットワークにおいては，データに接近する方法論は，多くの点で統計学と共通点を持っている．例えばパターン認識の目的は統計学の判別分析，クラスター分析と共通点を持っている．また学習理論における例題からの学習過程は，統計的推定論と同様にして展開される．しかし，それらの中には従来の統計学にはなかったアイディアが組み込まれている．温度・しきい値パラメータなど，いわゆるチューニング・パラメータはその典型である．最近，急速に展開されている独立成分解析，サポートベクターマシンなどについて，統計学者が参入して研究が行われている．

これらの方法論は，今後，多種多様な現実的課題に適用され，それらの解決に貢献していくことが期待されている．例えばベイズ推論は，構造の複雑な従属性をもつ現象をモデル化する場合に，それに柔軟に対処できる方法として発展すると思われる．また企業予測や病気の診断の際に見られるような，非線形でダイナミックな構造をもつ現象に対処するためには，非線形回帰やパターン認識，ニューラルネットワークなどの方法がますます発展すると予想される．さらにコンピュータ社会では，大量の情報が飛び交っているが，その中から真に有用な情報を抽出するデータマイニングにも，統計学で培われた方法が役に立つと期待できよう．環境のモニタリング，地下資源の探索などでは地理情報システムの構築が重要であるが，この分野での統計学の果たしている役割は今のところ小さい．画像処理や遺伝子解析，環境モデル化などの分野においても，統計学は使われ始めたばかりである．

情報化社会においては，”読み書きソロバン” の能力と同様に，一人一人が統計的・データ解析的な考え方をもつことがぜひとも必要になると思われる．すなわち ”統計リタラシー” は，コンピュータ・リタラシーとともに，現代人にとって必須のものとなるだろう．