

## 統計学とは何か C.R.ラオ

- ・統計学——不確実性を数量化し表現する方法
- ・数字にその素性を語らせるのは、統計家の熟練度や経験に依存するが、そのことが統計学が芸術であるとする所以
- ・記述統計学：位置や散布度の指標、さらには高次の積率や指標といった記述統計量によって、与えられたデータを要約し、度数分布、棒グラフ、2次元の図によってデータの際だった特徴を表示すること・・・記述的データ解析（DDA；Descriptive Data Analysis）
- ・推測統計学：データの要約は重視されるが、そこでは主に指定された確率分布の族（モデル）に関する議論が行われる。また、それらの標本分布は未知母数の推測における不確実性の幅を決めるためにしばしば利用される。・・・推測的データ解析（IDA；Inferential Data Analysis）
- ・カイ2乗統計量の導入により、与えられたデータが特定の確率分布族から発生したものであるという仮説の検定、すなわち、そのデータがある所与の仮説に一致するかについての検定が可能となった。
- ・推測的データ解析とは、ある特定の確率モデルの母数の推定、ある特定の仮説の検定、未来の観測値の予測、意思決定などの統計的方法論を指すものである。

ロシアの有名な確率論研究者 コルモゴロフ(A. N. Kolmogorov) 大数の法則

マハラノビス(P. C. Mahalanobis) インド統計研究所 (ISI) を創設

ラマヌジャン(S. Ramanujan)

映画「奇蹟がくれた数式」

statistics（統計学）という語の語源は、ラテン語で国家を意味する status であり、それは

「国家による、データの収集、処理、利用」を意味するために、18世紀の中頃にドイツの哲学者アッヘンヴァール（Gottfried Achenwall）によって作り出された新語である。

すでに述べたように、元々の語源的な意味では、統計学はデータの収集、編集、さらにデータを公的な政策作成へ適用することからなっている。

19世紀になってから、統計学はデータの解釈のための方法として、また意思決定のためにデータから情報を抽出する方法として、新しい意味をもち始めた。現在の傾向に基づいて、人間集団の社会的・経済的特性をいかにして予測することができるのだろうか？ 政府によって採用され

た、ある種の法制の影響はどのようなものであろうか？  
社会の福祉を推進するためには、どのようにして政策決定  
を行うのだろうか？ 明日は雨が降るだろうか？ 観測デ  
ータは、与えられた理論を支持するであろうか？ 収穫物  
の不作とか死などの破局的な結果に対して、保険契約を行  
うシステムを開発できるであろうか？

これらの種類の質問に答えるための最大の障害は、不確  
実性——原因と結果との間の1対1対応が欠如しているこ  
と——である。不確実性のもとでは、人はどのように行動  
するのだろうか。これは長い間人を悩ませてきた課題であ  
る。不確実性を懐柔することを学び、賢明な意思決定を行  
うための科学として**統計学が発展するようになったのは、  
20世紀の初めのことにすぎない。**

**演繹的推論によっては、前提の枠を越えるような新しい  
知識は創造できない。**なぜなら、導出されたすべての命題  
は、暗に公理に含まれているからである。

数学は“最高レベルの真実”とみなされているが、演繹  
法はその基礎をなすものである。しかしそれは、**必ずしも  
演繹法に論理的欠陥がないことを意味している訳ではな  
い。**

#### 2. 1. 4 アブダクション

時として、データに基づくことなしに、まったくの直観  
や想像力のひらめきによって新しい理論が生まれることが  
あるが、それは論理学の術語では、**アブダクション**とよば  
れている。それらは後に、実験を行うことによって確かめ  
られる。この有名な例としては、DNAの二重螺旋性、相  
対性理論、光の電磁気学的理論などが挙げられる。

帰納法とアブダクションとの違いは、かなり微妙であ  
る。帰納法では、実験データとそれに対して洞察を与える  
ための解析が、私たちの道しるべとなっている。しかし新  
しい知識の創造という究極の段階は、ある程度まで、以前  
の**経験や想像力の飛翔に依存している**。この意味で、帰納  
法はアブダクションであるとの信仰が生まれる。

以上をまとめると、**知識の進歩は、つぎの3種類の論理  
過程に依存している**と言える。

帰納法：観測データに基づく新しい知識の創

造.

アブダクション：データに基づくことのない、直観による新しい知識の創造.

演 繹 法：提案された定理の証明.

実際、事前分布を用い

ることなく推測の理論を発展させるための試みが、**近代統計学の創始者たち**である・K・ピアソン、フィッシャー、ネイマン・E・S・ピアソン・ワルドによって行われた†.

† K. Pearson (1857年3月27日－1936年4月27日), R・A・Fisher (1890年2月17日－1962年7月29日), J・Neyman (1894年4月16日－1981年8月2日), E. S. Pearson (1895年8月11日－1980年6月12日), A. Wald (1902年10月31日－1950年12月13日).

現在研究され、実際に適用されているような統計学は、科学なのであろうか、それとも技術または芸術なのであろうか？おそらく統計学は、それら三者を組み合わせたものである.

統計学は、固有のいくつかの基本的原理から導出された、幅広いレパートリーをもった技術に類似しているという意味では科学である。これらの技術は、決まりきった方法で使われるべきものではなく、与えられた状況に応じて正しい技術を選択するために必要な専門知識を身につけ、必要なら修正を行わなければならない。統計学は、ソフト・サイエンスにおいて経験的法則を確立する際に重要な役割を演じる。さらに、**統計学——不確実性を数量化し表現する方法——**の根拠に関しては哲学上の論点があるが、それは対象とする問題に応じて個別に論じられるべきものである。このように、広い意味では統計学は独立した学問であり、おそらく**すべての学問のための学問**なのであろう。

統計学は、産業生産における品質管理のプログラムのように、望まれる水準と安定性を維持するために操作システムのなかに、統計的方法論を組み入れることができるという意味においては、技術でもある。統計的方法はまた、不確実性をコントロールし、減少させ、それを考慮することによって、個人および社会の行動の効率を最大にするためにも用いることができる。

さらにまた、統計学は芸術でもある。なぜなら、帰納的推論に依存するその方法論は、十分には明文化されてはい

ないし、議論がない訳でもないからである。何人かの統計学者が同一のデータの組を解析したとしても、異なる結果に達する可能性もある。通常は、使用可能な統計的手段によって引き出されるものよりも、与えられたデータそのものの方が、より多くの情報をもっている。「レッドフォートの話」の例（第5章，2.14節参照）に見られるように、**数字にその素性を語らせるのは、統計家の熟練度や経験に依存するが、そのことが統計学が芸術であるとする所以である。**

これまで、記述統計学と推測統計学は異なった方法論をもつ2つの統計学の領域と考えられてきた。記述統計学の目的は、位置や散布度の指標、さらには高次の積率や指標といった記述統計量によって、**与えられたデータを要約し、度数分布、棒グラフ、2次元の図によってデータの間違った特徴を表示することであり、観測データを生成する確率的メカニズムや確率分布については何ら言及しない。**データを要約するこのような記述統計量は異なった種類のデータの比較などに利用された。分布の位置を示す平均値、中央値、最頻値といった指標のいずれを選択するかは、データの性質、回答すべき問題の性質などに依存し、その選択基準についてはいくつかの処方箋も用意されている。このような統計的解析は記述的データ解析（DDA ; Descriptive Data Analysis）とよばれる。一方、推測統計学においてもデータの要約は重視されるが、そこでは主に指定された確率分布の族（モデル）に関しての議論が行われる。データの要約統計量、あるいは記述統計量として何を選択するかについては指定された確率モデルに大きく依存する。また、それらの標本分布は未知母数の推測における不確実性の幅を決めるためにしばしば利用される。このような方法は推測的データ解析（IDA ; Inferential Data Analysis）とよばれる。

ピアソン（Karl Pearson）は記述的データ解析と推測的データ解析の溝を狭めようとした最初の統計学者である。ピアソンは、積率や度数分布といった記述的分析に基づく洞察から、確率分布族に関する推論を行うために**カイ2乗統計量とよばれる検定統計量を考案**した。カイ2乗統計量はピアソンによって定式化された最初の、そしておそらく最も有用な検定基準であろう。カイ2乗統計量の導入により、**与えられたデータが特定の確率分布族から発生したものであるという仮説の検定、すなわち、そのデータがある所与の仮説に一致するかについての検定が可能となり、新**

しい意思決定の方法の先がけとなった。

フィッシャーによって統計学の発展が推  
進されていたこの時代に、フィッシャー以外の研究者によ  
ってデータの背後に仮定される**確率モデルによらないノン  
パラメトリックな検定基準**が研究されていた (Pit-  
man [26])

初期の統計学の発展期においては、統計学の研究の多く  
は生物学の問題に由来したものであったが、それと並行し  
て、小規模ながら工業生産における統計学の利用の研究も  
発展した。シューハート (Shewhart [31]) は、管理図の利用  
によって生産過程における変化を検出するグラフ的手法を  
開発した。この方法は外れ値や変化点の見つけ方について  
の最初の研究であろう。

ワルドは**抜き取り検査** (sampling in-  
spection) に適用可能な逐次的方法を開発した

データ解析の図式

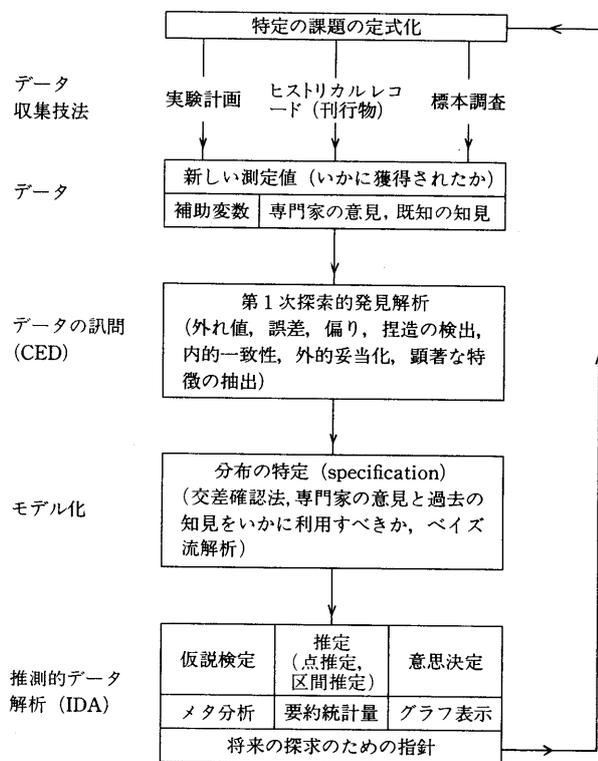


図 3.1 実生活における課題

データの訊問 (CED ; Cross Examination of Data) とは、  
データの性質を理解し、測定誤差、記録誤差や外れ値を検

出し、事前情報の妥当性を検証し、データが捏造されたものか否かを検証するために行われる探索的または第1次的研究である。

多変量のデータにおいて記録された値や測定値の誤りの個数が多くない場合には、個々の測定値や比の値に関する度数分布、2組の測定値の対の2次元散布図などを描いたり、各測定値の最初の4つの積率、および歪度、尖度の指標 $\gamma_1$ 、 $\gamma_2$ を求めることによって記録誤差や測定誤差を見いだすことが可能である。とくに、歪度、尖度は外れ値に対して鋭敏である。

ここでは、データに含まれる外れ値や、矛盾のあるデータをどう扱うかといったやっかいな問題にどのように対処すべきかについて考えてみよう。

極端な値、すなわち、ある意味で他の値と整合性をもたか、値をどのように取り扱うべきであろうか？ 外れ値やデータの汚染に関する問題はやっかいなもので、今日の研究課題の1つである。不幸なことに、データの合理化（rationalization）やデータのトリミングのように、データに何らかの統計的補正をすること以外には上記の課題に対する満足すべき方法は存在しない。

“お蔵入り問題、file drawer problem

統計的に有意であるとの結論が得られなかった論文は没。

推測的データ解析とは、ある特定の確率モデルのものと未知母数の推定、ある特定の仮説の検定、未来の観測値の予測、意思決定などの統計的方法論を指すものである。

統計解析の目的は、観測データからすべての情報を抽出することである。記録されたデータは記録誤差や外れ値のような欠陥をもつことがあり、時には捏造されている可能性もある。統計家が最初になすべきことは、データの訊問や精査を行ってデータの欠陥を発見し、データの特徴を理解することである。つぎの段階においては、データに関する事前情報や交差確認法の技法を利用することによって、与えられたデータに適

切な確率モデルを想定することである。こうして選ばれたモデルに基づいて**推測的分析**，すなわち，**未知母数の推定**，**仮説検定**，**未来の観測値の予測**，**意思決定**などが行われる。データをいくつかの異なったモデルのもとで検証することは，想定されるいくつかのモデルに対して頑健な方法を用いるよりも，より多くの情報を生み出すことになる。最後に強調しておきたい点は，データ解析は新しい課題を生み出し，将来の研究計画に役立つ情報を提供するに違いないということであろう。

切断・・・下記は理解できない。

一般に，確率変数の確率密度関数（連続変数の場合は確率密度関数，離散変数の場合は確率）を  $p(x, \theta)$  とする。こ

こで， $\theta$  は未知パラメータである。また，確率変数  $X$  のとる値は標本空間  $\Omega$  の部分空間  $T(\subset \Omega)$  に制限されているとする。このとき，切断確率変数  $X^T$  の確率密度関数は

$$p^T(x, \theta) = \frac{w(x, T)p(x, \theta)}{u(T, \theta)} \quad (2.1)$$

である。ここで， $w(x, T)=1$  ( $x \in T$  のとき)， $0$  ( $x \notin T$ )，および  $u(T, \theta)=E[w(X, T)]$  である<sup>[訳註 24]</sup>。式 (2.1) は元の確率密度関数を適当な関数で重み付けたものであって，重み付き分布の1つの例になっている。重み付き分布の一般的定義は次節で与えられる。

この意味で，統計的知識は国の資源とみなされている。

近代の発見をテーマとした最近の書物において，**統計的品質管理を20世紀における偉大な技術的発見の1つ**としてとりあげていることは驚くにあたらない。実際，統計的品質管理のような技術的発見はめったに生まれるものではない。**統計的品質管理の理論は非常に単純**ではあるが，その適用可能性は広く，しかも容易に利用可能である。しかしながら，得られた結果は効果的であり，少額の投資でその見返りは大きなものとなる。

シェイクスピアの大多数の作品の発行日付は記録に残っているが，記録が残っていない作品もある。発行日付がわかっている作品を用いて発行日がわかっていか他の作品の発行日をどのようにして推定できるであろうか？ ヤルデイ (Yardi [19]) は作品に関する外的な資料を集めることなく，純粹に定量的手法のみを用いることによってこの問題の解決に取り組んだ劇作品のそれぞれについて，(i)

冗長な最後の音節, (ii) 完全に分かれている行, (iii) 休止 (ポーズ) だけをもち分かれていない行, (iv) 会話文の総数, のそれぞれについての頻度を求めた. このようにして文学的スタイルを数量化することにより, ヤルデイは発表の日付が既知の劇作品に関するデータを用いて, シェイクスピアの文学作品中の長年にわたる文体の変化を分析した. こうして, **補間法を用いる**ことにより, 作品『間違いの喜劇』の発表時期は 1591~92 年の冬, 作品『恋の骨折り損』の発表時期は 1591~92 年の春と推定された.

ココヤシの木では, 実は茎にらせん状についている (らせん葉序という) が, そのらせんの方向によって左巻きや右巻きの分類が可能であることは一般的には知られていない.

...

なぜ, ある木は左巻きで, 他の木は右巻きなのであろうか? この性質は遺伝するものであろうか?

...

そこで調査を行ったところ, 北半球における標本から得られた左巻きの割合は 0. 515, 南半球からの標本における左巻きの割合は 0. 473 ということが判明した. この差はおそらく地球が一方向に回転することの影響によるものであろうと考えられる. この原理は浴槽の渦現象 (栓をぬいて浴槽から水が流出する際, 渦が左または右のいずれかにまく現象) の説明にもなっている. すなわち, 巧く管理された状態のもとでは, 渦は北半球では反時計方向, 南半球では時計方向となる.

もし, デーヴィスが左巻きと右巻きの木において異なる何らかの特徴を探求することにこれ以上熱心でなかったならば, 研究はいくらか学問的性格を帯びたものという程度にとどまっていたであろう. **彼は 12 年間以上にわたって, 農園において左巻きと右巻きの木の収量の平均値を比較し, 左巻きの木の方が 10% も多くの収量を上げることを発見して仰天した.** この結果に対して何らの説明もできなかったとはいえ (この疑問は追求されなければならないが, それほど簡単に解決できるものではないであろう), この経験的に与えられた結論というものは経済的に多大な重要性をもつものである. 左巻きの木のみを選んで栽培することにより, その収量が 10% も増加するとは一体どういうことであろう!

## 証明

の程度に関する基準はつぎの用語を用いて表されている：

- (1) 証拠の優越 (the preponderance of the evidence)
- (2) 明白かつ確信的な証拠 (clear and convincing evidence)
- (3) 明白であいまいでなく、かつ確信的な証拠 (clear, unequivocal and convincing evidence)
- (4) もっともな疑いのない証明 (proof beyond reasonable doubt)

これに対して、決定論的方法によって生成された数は、局所的にはランダムのようなふるまいをするが、大局的にはある規則性をもつことが観察される。ここ 20 年の間に科学者たちは、この後者のタイプの現象を、“カオス”という名をつけて研究を開始しだした。雲の形成、乱気流、国の海岸線などのような複雑な輪郭や形状をモデル化したり、価格の変動をも簡単な数学の方程式を用いて説明するために、新しい接近法が提案されている。この種の考え方は、あるシステムの出力結果を記述するのに偶然性のメカニズムによるのとはいささか異なっている。偶然性は無秩序中の規則性を論じるのに対して、カオスは、秩序中の無規則性を論ずるのである。この両者とも観察された現象をモデル化するのに適用できる。

カオスについての研究は、ローレンツによる“バタフライ効果”とよばれる現象の発見とともに注目されるようになったが、それはあるシステムが初期状態に鋭敏に依存するような効果である。彼は長期の天気予報においては、予測公式の入力として使われる初期の測定値におけるわずかな誤差が、予測値においては非常に大きな誤差にまで拡大してしまう可能性のあることに気がついた。

人生を興味深いものとする 2 つの基本的な要素は、偶然性とあいまいさのように思える。それは自然事象の予測不可能性であり、私たちが意思の疎通を行う際に使用する語が一意的に解釈できにくいことである。

## 興

興味深いのは、 $\pi$  の十進数表示を求めるのに当たって、次の神秘的なラマヌジャンの公式の変形版を用いた計算がコン

コンピュータで行われているということである。

$$\frac{1}{\pi} = 2\sqrt{2} \sum_{n=0}^{\infty} \frac{\left(\frac{1}{4}\right)_n \left(\frac{1}{2}\right)_n \left(\frac{1}{3}\right)_n}{(1)_n (1)_n n!} (1103 + 26390n) \left(\frac{1}{99}\right)^{4n+2}$$

ここで、 $(a)_k = a(a+1)\cdots(a+k-1)$ である。) しかしなが

人は優れたアイデアをどのようにして得るのであるか？ 創造的になるためにはどんな準備が必要なのであるか？ 天才は生まれながらのものなのか、それとも作られるものなのだろうか？ これらの疑問に対して明確な答はない。しかしながら、答が見つかったとしても、このように輝かしいアイデアがラマヌジャンの頭のなかからどうして瞬時にひらめいたかは説明できないであろう。それよりも興味をそそることは、**彼は高度な数学について正規の教育を受けていなかったため、数学的研究の手ほどきを受けておらず、また、現代数学の研究分野や方向にも気づいてもいなかった**ということである。彼は証明なしで定理を述べ、その動機づけが何であったかも述べてはいない。ラマヌジャンはどのようにしてこれらの結果を得たかを説明することはできなかったのであろう。彼はナマッカルの神々が夢の中で自分に靈感を与えてくれたといった。ベッドから起きて、いくつかの結果を書き留め、必ずしも厳密な証明を与えることはできなかったにしても、即座に証明したこともたびたびであったろう。ラマヌジャンによって述べられた定理の多くは正しいことが示されている。創造性は潜在意識のレベルで起こるものなのであるか。

マハラノビス教授は、ラマヌジャンと同時期に英国ケンブリッジ大学に在籍していた。彼はラマヌジャンに関連したいくつかのエピソードを話している。

<訳者あとがき>

ここで、本書の内容を概観してみよう。まず、第1, 2章においては、現実世界において私たちがしばしば遭遇する不確実性について言及し、私たちの日常生活における不確実性を考慮した意思決定の方法、およびそれに辿りつくまでの歴史的発展が解説されている。第3, 4章においては不確実性に対処するための統計的データ解析の方策が展開されている。とくに、外れ値や記録ミスを発見するためにデータの訊問や精査を行うことの必要性、および適切なモ

デル設定の必要性が強調されている。後者については、データがどのようにして得られたかに応じて、データの調整を行う必要性を論じ、その1つの方法として、重み付き分布に焦点をあてて解説している。第5章では、さまざまな研究領域において、統計的方法による新しい知識の獲得、および最適な決定がいかに行われるかについて、18の具体例を通して解説している。第5章で取り上げられている統計的手法は比較的平易なものであるが、使い方によっては、平易な手法が現実世界における困難な課題を解決に導くことがある、という主張は斬新である。第6章では統計学がいかに社会的理解を得ることができるかという点に焦点をしばりながら、統計学の利用の望ましいありかたについて論じている。