書　評

影浦　峡（著）

計量情報学
～図書館／言語研究への応用～

丸善株式会社，2000年，2,600円（税別）
ISBN4-621-04758-2

サイコロを6回振ったとき，4，2，6，4，1，1という目が出たとしよう．このとき，将来においてどのような目が出るかを予測するのは難しい問題である．この標本データにおける目の数の平均が3だということはすぐに分かるが，注目したいのは，まだ3と5が出ていない，すなわちすべての事象が出揃っていないという点である．

サイコロだともう少し振る回数を重ねればすべての事象は現れそうだ．しかし，もっと複雑な問題，たとえば，ある学問分野における（潜在的な用語も含め）全専門用語の6割をカバーする用語辞書を作る場合はどうだろうか．仮に1000論文調べて2000用語が集まったとしても，これだけで全専門用語の6割が出揃っているかどうかを判断することは大変に難しい．

著者の近著「計量情報学」はこのような問題を正面から取り上げた貴重な1冊である．著者は，上で述べたような専門分野の語彙の増大（語彙成長）の定量的な分析を研究しておられるが，この本では，より分かりやすいモデルとして，ある分野における著者と論文数の関係を用いて議論を進めている．

以下に章ごとの内容を簡単に述べよう．

第1章では，まず壺モデルによる2項分布について解説されている．そして，本書の主題である頻度統計についての導入的な説明がなされている．頻度統計とは，標本中における「異なり事象」とその各々の頻度についての統計であり，人間の身長のように母集団に連続的確率分布を想定する場合とは異なる分析が必要である．つまり，母集団の事象のうちどれだけがすでに標本中に現れているかを問題にしなくてはならない．

第2章から**第4章**では，統計量の標本数への依存性を解析することの重要性について述べられている．特にある標本量のデータが得られたとき，そのデータだけを対象にするのではなく，標本量が変化したとき統計量がどのように変化するかに関するモデルを考えることの重要性が説かれている．そして，標本量に依存しない統計量の有力な候補としてSimpsonのDという不偏推定量が示される．

ここでDについて簡単に説明しておこう．たとえば，(a) 9人の著者が1本ずつ論文を書いている場合と，(b) 9人の著者が2本ずつ論文を書いている場合を比べる．このとき，より多くの論文の著者を調べた場合に，もう新しい著者が現れない可能性は (b) の方が大きいと予想される．Dとは，このような可能性を表す指標である．

後半の**第5章**から**第7章**は標本量の変化に応じて未出現事象の頻度をどのように見積もるかという問題を解説している．この問題は，確率的言語処理においてコーパスに出現していないN-gramや語彙などの出現確率を見積もるという工学的問題に直接応用できる．つまり，確率的言語処理では，既存のコーパスに出現しないN-gramや語彙が将来どのくらいの確率で出現するかを見積もっておかないと，未知の事象に十分対応できない．この問題に対して，確率的言語処理ではGood-Turingの見積もりを用いるが，他書ではあまり触れられていないこの見積もりの導入を**第5章**で丁寧に説明している．

さて，**第5章**から**第7章**での本題は，与えられた標本量より小さいか，あるいは大きい標本量に対応する特定の頻度の異なり事象数の期待値を計算するモデルである．具体的には**第6章**で2項補間・補外について述べられており，**第7章**で拡張Zipf族関数を用いた補間・補外に展開している．これらについて，いくつかの学会の著者とその著作論文数との関係を収集した実際のデータを用いた評価も示している．その結果，実際のデータよりも小さい標本量の場合は非常に正確な予測ができる．また，実際のデータより大きい標本量の場合も与えられた標本量の2倍程度までは，小さい標本量の場合よりずれはあるものの，かなりよい予測ができることが示されている．なお，この後半の内容については，この本が数学的側面について拠り所にしているBaayenの著書の翻訳も同じく著者によって進行中であり，その刊行が期待される．

評者は言語情報処理の観点から用語や対訳辞書の自動抽出の研究をしているが，この本からは大変役立つ情報を知ることができた．この本は大量の数式を含むので，一見して文系の読者には難解な感じを与えるかもしれない．しかし著者は，数式を示すだけではなく実際のデータも併用して，数式の表す具体的な意味を分かりやすく解説している．この解説は研究上の示唆も多く含んでいる．情報にかかわる諸学問ないし技術分野の学生，技術者，研究者であれば，文系，理系を問わず一読をお勧めしたい．

（中川裕志／東京大学情報基盤センター）