第 1 章

統計学とは何だろうか？

1.1 統計学の歴史

統計学の歴史については，さまざまなことが言われている。ラオ（1993）によれば，統計学の歴史をずっと過去に遡っていくと，本来は，「家畜や他の財産の帳簿をつけるために原始人が木につけた刻み目」だという[*1]。その後，国家を経営するための基礎資料的な意味あいが強くなった。ラオ（1993）には，「ある国およびそこに生きている生命の状態や発展についての，もっとも完全で，もっとも根拠のある知識」という，マルシャスの言葉が引用されている。

英語の statistics は，ラテン語で国家を意味する status を語源として，18 世紀半ばにドイツの哲学者アッヘンウォールが作った言葉が元になっている（ラオ，1993 年）。国家としての人口規模が大きくなるとともに，雑然とした大量の生データを，解釈をやさしくしたり種々の方策決定に用いるためにまとめ上げる手法が必要になり，グラントの生命表やケトレーの度数分布図など質的にも量的にも統計手法が開発されてきた。産業革命が進行する中，1834 年，英国王立統計協会設立により，学問としての「統計学」が成立し，「人間に関係することがらで，数量で表現することが可能で，一般的な法則を導き出すのに十分なだけ積み重ねられたもの」と定義された。

20 世紀末，コンピュータの進歩とともに統計学の理論や技術も大きく進展し，同時にパッケージソフトウェアの開発によって，統計学の専門家でなくても統計解析を行うことが可能になった。そのため，統計学は，いまやすべての自然科学や社会科学で適用される科学的分析の技術となっている。もっとも広い意味で定義するならば，「不確実性を考慮した論理的推論」ということになるだろう。

[*1] 「原始人」という呼び方には問題があるが，大事なことは，文字が無かった時代であっても統計的な概念は必要だったし，ありえたということである。

1.2 不確実性とランダム（乱雑さ）

世の中のほぼすべての事象は不確実性を含んでいる。素粒子レベルでは物理法則も不確実性を含むし（ある原子核に含まれる電子が存在する確率がゼロでない場所という意味で電子雲は決まるけれども，電子がある瞬間にどこに存在するかということは確率的にしかいえない），遺伝子の発現や社会における個人の行動なども，決して決定されてはいない。

不確実性を数学的に扱うには，確率的に起こるできごと（確率事象）を扱わねばならない。確率事象は，一般に，何回中何回くらい起こりそうかはわかっているが，いつ起こるかがわからない。そのために，ふつうは既知の分布関数が使われる。ただし，コンピュータ上で扱うときは，ランダムな数字の列，すなわち乱数列[*2]を利用することができる。たとえば区間 (0,1) の実数値をとる一様乱数列[*3]を確率事象に割り当てれば，その値が確率 p より小さいときに事象が起こり，確率 p より大きいときに事象が起こらないと解釈することによって，確率を事象が起こるか起こらないかに置き換えることができる。このやり方で確率分布をシミュレートすることは，コンピュータ集約型統計学と呼ばれる分野で近年盛んに行われている。

乱数列については，線型合同法など，数式を使って生成される擬似乱数列というものがあり，数式がわかれば次の数字は予想できるのだが，見かけ上はでたらめな数の並びに見え，実用上十分なでたらめさをもっているので，コンピュータ上で乱数列が必要な場合は良く使われている。現在ではコンピュータ上でも熱拡散の状態を測って真の乱数を得る拡張ボードが市販されていて（たとえば東芝のランダムマスターなど）利用可能だが，メルセンヌツイスター[*4]のような優れたアルゴリズムで生成された擬似乱数を使う方が普通である。

[*2] 次の数字が予想できない，意味のないでたらめな数の集まり。たとえば，20 桁の対数表の 15〜19 桁目を並べたものとか，袋に入れた 500 個ずつの白ビーズと黒ビーズから，よく混ぜて 1 個ビーズを取り出して色を記録し，元の袋の中に戻して，取り出す前と同じ状態に戻してからまたビーズを取り出して色を記録し，と繰り返した（復元抽出した）ときの色の列など。

[*3] 乱数列のうち，各数字の出現頻度が等しいと期待されるものを一様乱数列と呼ぶ。

[*4] 松本眞，西村拓士両氏によって，1996 年から 1997 年にわたって開発された擬似乱数生成アルゴリズムで，生成速度が速く，きわめて周期が長く，C で書かれたプログラムソースコードが自由に利用できるという利点をもつ。詳細は `http://www.math.keio.ac.jp/home2/matumoto/public_html/mt.html` を参照されたい。

1.3 統計解析の手順

データの分析技術としての統計解析は，一定の手順を踏んで行われる。箇条書きすると，以下のような手順が典型的と思われる。

- 目的を明確にする
- 生データをとる
- データ化（エディティング，コーディング，データ入力）
- データの図示（幹葉表示やヒストグラムなど）
- 代表値（分布の位置やばらつきを示す値）の計算
- 作業仮説の明確化（ここで因果関係についての仮説を立てることが多い）[*5]
- 仮説検定や区間推定を行う（攪乱要因に配慮し，その影響を制御する必要がある）
- 因果関係についての推論を行う（先行研究の知見なども総合する必要がある）

データ化以前の段階についての細かい説明は調査法についての本やデザインの本を参照されたいが，大事なことは，データを取る前の段階で，統計解析をどうするか考えておくべきだということである。実際にはなかなかそうはできず，データを取った後で解析法が考えられる場合が多いのだが，後付けの分析はバイアスの元になるし，言いたいことを検討するための解析に必要なデータが取れていないことが解析段階で判明しても，後の祭りなのである。だから，統計解析は，データを取ったあとで始まるものではなく，データを取る前の段階で始まっていることを肝に銘じておくべきである。フィールドワーカーや実験科学者や政策担当者は，データを取ってから解析法に困って統計学者に相談するのではなく，デザインの段階から相談すべきである[*6]。

因果関係については，数値間に常に関連があるというだけではなく，時間的前後関係など，いくつかの条件を満たさないと因果関係があるとはいえないし，その条件についてもいろいろ議論がある（Rothman, 2002）。一般には，図 1.1 に示すように，その分析で着目している結果（英語では outcome と呼ばれる。この図の場合なら高血圧）を評価軸としたとき，「A：この分析で結果との関係を評価したい因子（この図では体脂肪割合や食塩摂取量）」，「B：この分析で結果との関係をみたいわけでは

[*5] 因果関係については，佐藤・松山 (2002) の議論がすばらしく良くまとまっているし，Rothman (2002) の第 2 章の議論も参考になる。後者は
http://www.oup-usa.org/sc/0195135547/media/0195135547_ch2.pdf
として web で全文が公開されている。

[*6] または，フィールドワーカーや実験科学者といえども，きちんと統計学を学ぶべきである。

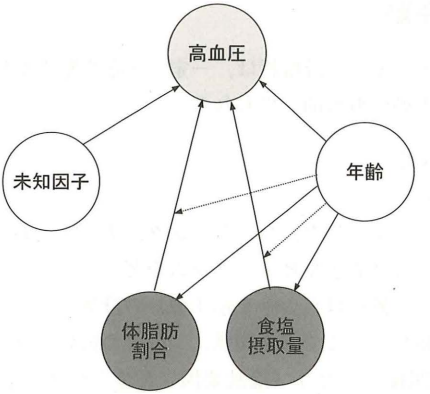

図 1.1 因果関係と攪乱要因の例

ないが結果やAやそれらの関係に影響することがわかっている因子（この図では年齢)」,「C：この分析では調べていないが，結果に影響する因子（この図では未知因子)」に分けて捉えることができる。Bは攪乱要因とか交絡要因と呼ばれ，Bの影響を調整した上で，Aと結果の関係を調べないと，真の関係はわからない。結果のうち，Aによって説明できない部分が偶然変動及びCとして残るので，Cの影響は小さいほどよい。その意味で，統計解析では結果のどれくらいがAによって説明できるかを調べたり，Cの影響がどれくらいあるかを調べる手法があり，因果関係の確からしさを判断する上で重要である。

1.4 統計解析の2大方針

上に述べた意味での手順は同じだが，推論の根拠について考えたとき，統計解析には，大きく分けて2つの異なる方針がある*7。

1つは，デザインに基づく解析である。仮説検定の例としては，並べかえ検定やログランク検定がこれに当たる。デザインに基づく解析では，データが，ランダムにデータを取った場合に得られるパタンの1つであると考え，その確率を直接計算す

*7 この考え方について詳しくは，松山裕「統計解析の2つの原理」，帝京大学研究用コンピュータ室ニュース No.25，pp.54-64，1992年7月31日を参照されたい。

る。攪乱要因はデータを層別することで制御する[*8]。一般に計算量は多くなるが，分布を仮定する必要がないのが利点である。

もう1つは，モデルに基づいた解析である。t 検定や重回帰分析や比例ハザードモデルなど，有名な多くの統計解析は，この考え方に基づいている。デザインに基づく解析に比べると，一般に計算量は少なくて済む。結果の分布を記述するために確率分布を仮定し，その未知パラメータをデータから推定する。データがモデルにもっとも良く当てはまるようなパラメータを最小二乗法や最尤法などで推定し，当てはまりの悪さを AIC などの指標を使って評価する。攪乱要因の影響は，説明変数としてモデルに入れることで調整することになる。

1.5 統計解析の道具

実際の統計解析は，コンピュータのソフトを使って行われるのが普通である。SAS と SPSS がもっとも有名でよく使われているソフトだし高機能だが高価である。会社などで十分な予算があれば，SAS を使ってもよいであろう。しかし個人が使うには非現実的な値段である。SAS インスティテュートが販売している JMP はそれほど高価ではなく，マウスでデータを見ながら操作できるので取っ付きやすいが，複雑な操作を一度に実行させたり，大量の計算をさせるにはあまり向いていないように思われる（ただし，バージョン 4 以降はスクリプト言語をサポートしたので，R ほど柔軟ではないと思うけれども，複雑な操作を予めスクリプトファイルに書いておくという使い方も可能になった）。

また，世間では，Microsoft Excel を使って初歩的な統計解析をすることも多いようである。覚えておくといろいろ便利かもしれないが，解析の中身がブラックボックスだし，それがないと何も出来ないという状況では困る。そもそも統計ソフトではないので，不正確な結果がでることもある[*9]。少量のデータ入力や変換には便利（大量のデータ入力にはデータベースソフトを使うか，html のフォームと cgi を組み合わせて入力環境を作ると良い）だが，Microsoft Excel のマクロ言語などで統計解析をすると，後で何をしたのかわからなくなりやすいし（その点はログを取れないメニュー式のソフトはすべて同じ危険を孕んでいる），ちょっとだけ修正するといったことがやりにくいので，本格的に統計データ解析をするには薦められない。

本書では R を利用した解析方法を説明する。R の最大の利点は，オープンソースであり，かつ拡張性が高い点だと思うが，慣れれば使いやすさもかなり高い水準にあ

[*8] 攪乱要因の値が異なる対象ごとに別々に分析することを層別解析するという。

[*9] 詳細は `http://aoki2.si.gunma-u.ac.jp/Hanasi/excel/index.html` を参照されたい。

る。少なくとも表計算ソフトや汎用言語を使うよりは，ずっと簡単に統計処理ができる。R についての詳細は，付録を参照されたい。