〔第2回〕

統計解析の基本的考え方

新田裕史*　高木廣文**

はじめに

数式を書き並べ，「統計学とは何か」というようなことを冒頭から提示することは，賢明な方法ではないと思われる。かえって，「統計アレルギー症」の患者を増やすだけであろう。統計学の専門家ならば，そのような議論も有用かもしれないが，「統計学で何ができるか，何がわかるか」を理解することの方が，医学に携わるものにとっては必要なことである。統計解析の基本的な考え方を理解し，正しい使い方の基本を身につけ，さらには統計的な方法を用いて導かれた結論の持つ性質を見極めることが求められている。本シリーズでは今回以降，このような観点から統計学について解説していきたい。

今回は，統計的なものの考え方の最も基礎的な部分について解説したい。

I. 統計データとは

統計学で扱うデータはすべて，測定ないし観察されたものである。統計学の基本は，得られた測定値や観測値には「誤差」が含まれているという前提に立つことである。例えば，ある患者の血清IgE総量の測定を繰り返し行ったとしよう。それらの測定値は一定ではなく，あるバラツキを持っているのが普通である。このような場合，統計学では測定値が以下のように，

1）〔測定値〕＝〔真の値〕＋〔誤差〕

と，真の値に何か誤差が加わって，バラツキが生じていると考えるのである。もちろん，真の値も各種の状況下で変化するかもしれない。ただし，その変動の仕方には規則性があると仮定する。ところで，血清IgEを同じように測定した場合でも，複数の患者について測定を繰り返した場合には，1）の場合とやや状況が異なる。基本的な考え方に変わりはないが，この場合には，

2）〔測定値〕＝〔集団としての真の値〕＋
〔個人としての真の値〕＋
〔誤差〕

のように真の値を2つの変動要因に分けて考える方が，より一般性があり，妥当かもしれない。このように，データの構造をモデル化して解析することは統計解析では非常に重要なことである。

われわれは何か測定値が得られたときに，データ全体の平均値を求めたり，グループ別の平均値を求めたり，さらに個人ごとの平均値を求めたりしている。これは，仮定されたモデルを通して真の状態を探ろうとしている作業なのである。

測定値や観測値という量を示す数値のみを考えがちである。しかし，実際の統計データには質的なものも多数含まれている。例えば，患者の性別や特定の症状の有無などである。また，治療効果を，「著効」「有効」「やや有効」「不変」「悪化」の5段階で評価したものも質的データである。

統計データの種類によって，仮定される統計モデルが異なってくる。統計データの種類が同じでも，データの取られ方，言い替えれば研究のデザインが異なれば，仮定される統計モデルも違ってくる。ここでは，まず統計データの種類について

* 東京大学医学部保健学科疫学
〔〒113 東京都文京区本郷 7-3-1〕
** 文部省統計数理研究所

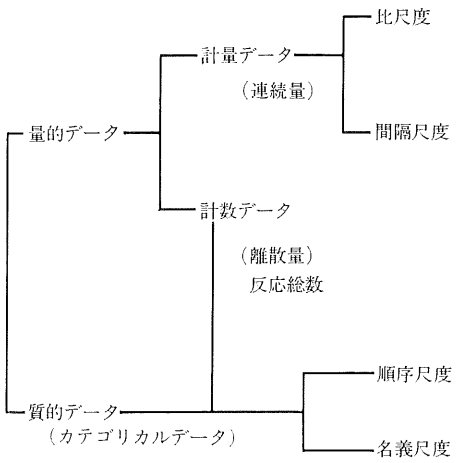

図 1 統計データ種類

解説する。

統計で扱うデータには大別して量的データと質的データがある。質的データは離散量データと連続量データに区別できる(**図 1**)。それぞれ，計数データと計量データと呼ぶこともある。計数データは白血球数などの，ゼロまたは正の整数値をとるものである。計量データは物質の温度や血液中のある物質の濃度など連続的な尺度で測られるデータである。計量データのうち，温度のように原点 0 に絶対的意味がなく，値の差だけに意味があるものを間隔尺度という。これに対して，物の長さのように原点に具体的な意味があり，値の差だけではなく，比にも意味があるものを比尺度という。ある疾患の患者数などの質的なデータの特定のカテゴリに対する反応総数もまた計数データとして扱われる。計数データは基本的に比尺度の性格を持っている。

質的データは性別や病気の有無（分類の細目をカテゴリと呼ぶ），学科の成績など質的なカテゴリに対する反応を表すものである。質的データのうち，性別における男女のようにカテゴリが名義的なものと，科目の成績を A, B, C, D で表したときのように，その順序に意味がある場合がある。前者を名義尺度，後者を順序尺度と呼んでいる。なお，質的データをカテゴリカル・データと呼ぶこともある。実際の解析においては，質的データについても，それぞれのカテゴリに数値を与えておく場合も多い。例えば，性別で「男」には 1，「女」には 2 を与えるような場合である。

表 統計データの形

	変数 1	変数 2	変数 3	……
個体 1				
個体 2				
個体 3		＜観測値＞		
⋮				

このような統計データは通常，**表**のような形式で作られる。

データを収集する場合でも解析を行う場合でも，「個体」は測定の単位である。人を対象としていれば各個人であり，物の分析をしているのならば各サンプルである。「変数」は測定・観測の項目であり，これまで述べたようにさまざまな尺度が含まれているのが普通である。対象者の性・年齢などの属性，生化学的検査値，アンケート調査の集計データならばそれぞれの質問項目への回答である。同じ対象者で何項目かの測定をしていれば，それぞれの測定値が「変数」である。「個体」のことを「ケース」や「オブザベーション」と呼ぶこともある。

観測の単位となる，もしくは単位とみなすものに関する観測値が複数ある場合には，各観測値を 1 行に横並びにすることが原則である。そして，同一の観測項目は同一の列に並ぶことになる。時系列的に観測を続けた場合のデータのように，状況によっては「個体」と「変数」の区別がつきにくいこともある。

II. 母集団と標本

統計解析の目的は，標本に基づいて母集団に関する推測を行うことであるといわれる。母集団とは，統計解析の対象となる「同一の基本的概念を共有する特定の特性を持った個体の集合体」のことである。母集団は，統計学では極めて抽象的な概念上のものとして設定されるが，実際の調査では実体のある母集団が設定される。この母集団か

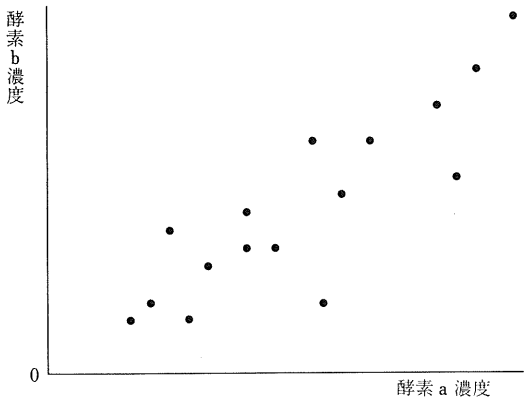

図 2　aとbの濃度の相関図

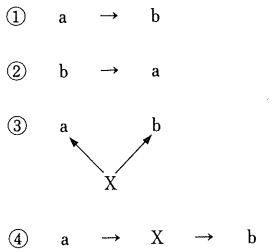

図 3　2つの変数の関連性のタイプ

ら，一定の方法によって選ばれて，測定や観察の対象となるものを標本と呼んでいる。

例えば，ある疾病の患者を対象として何かを調べようとする。今，その疾病の患者が50人いるとしよう。統計解析の対象はその50人のように思われるが，実は対象となるのは「その疾病を持った人」の母集団である。すなわち，50人の患者のデータを解析して得た結論は，その50人にのみ当てはまることではなく，過去，現在，未来にわたり，無限に存在するであろう「その疾病を持った人」の集合についての結論を導きたいのがふつうである。このシリーズの第1回で強調されていたように，それが「普遍性」を持つ結論ということである。

統計解析の対象となる母集団は常に，「ある特性を持つ」ものの全体というような仮説的な，無限の数を持つものとは限らない。選挙結果の予測のための世論調査の対象集団は有限の実体のある母集団である。しかし，医学の分野では仮説的な無限の母集団を対象と考えた方が有用と思われる。

われわれは標本に基づいて母集団のことを推測しようとする。これを統計的推測と呼んでいる。これは一部から全体を推測しようとするものであり，当然誤りを含むものである。統計解析ではこの誤りの程度を確率を用いて表現する。この点で，統計的推測は臨床的な判断と相容れない一面を持っている。それは「集団」と「個」の問題と言い替えてもよいだろう。統計的推測は集団に対する判断を行うものであり，集団の構成員である「個」に対する判断を行うものではない。統計的推測の結果として出てきた集団における情報が，「個」に対してどのように還元されるかは非常にむずかしい問題である。

ところで，統計的推測だけが統計解析の仕事ではない。標本の持つ情報によって母集団の特性を推測するためには，まず標本それ自身の情報を整理して記述する必要がある。この仕事を，記述統計と呼んでいる。つまり，統計的推測に先だって，記述統計の段階が存在する。統計的推測の手法はデータに対してさまざまな仮定を置いている。このような仮定が満たされているかどうかを判断するためにも，記述統計が必要である。その意味では統計的推測のための予備的解析と位置づけることができるだろう。

III. 統計的関連性と因果関係

統計的推測の結果，ある一定の信頼度のもとで，ある事象AとBとに関連性が認められたとしよう。例えば，血清中に存在する2つの酵素濃度の関係を描くと**図2**のようになったとしよう。統計学では，これをaとbには相関関係が認められたと表現する。このことは，「aの濃度が高いときには，bの濃度も高い」などといい表わされる。しかし，このことを「aの濃度が高くなれば(その結果として)，bの濃度も高くなる」ことと誤解してはならない。

aとbに相関関係が認められたという場合，aとbの真の関係にはいくつかのタイプが想定できる。ひとつは，その関係が偶然生じた場合であ

る。統計的推測によって，この可能性はある一定の確率以下になることが保証できる。関係が偶然生じたものでない場合でも，いくつかの異なるタイプがある（**図3**）。①と②は因果関係を想定できる場合であるが，統計解析だけでは因果の方向を推測できない。また，③の場合は，第3の変数Xがあり，aとX，bとXにはそれぞれ因果関係が想定できるが，aとbは直接関係がなく，見かけ上相関関係がみられる場合である。第3の変数が④の場合のように，aとbの中間に位置している場合も有り得る。

相関関係に限らず，統計的関連性は必ずしも因果関係を示していないことを，肝に銘じておく必要がある。

次号では，平均や標準偏差など記述統計でよく用いられる指標を中心に解説したい。

*　　*　　*