統計学の基礎概念

1　統計への序章

統計、とうけい　とは何だろう。イギリスのグリニッジ天文台を通る経度をゼロ度として、それから東に測った経度でもないし、鶏を囲いの中にとじ込めて戦わせて、それを観戦するという残酷な遊びでもない。ＳＴＡＴＩＳＴＩＣＳ である。言葉の語源は，ものごとの状態を表す意味にもちいられる，ラテン語の status から由来するといわれる。現代の統計学は１９世紀後半から主にイギリスで発達した数理統計学とドイツで発達した社会統計学が知られている。たんに統計学というと前者の数理統計学をさすことが多い。われわれのまわりをちょっと見回しても、そこには多くの統計データ、たとえば、市町村の人口動態、降雨確率、物価指数、テレビの視聴率、株式市況、学力試験の偏差値などに気付かれるだろう。この漠然と理解されていて、日常生活に結びついた重要な概念をこれから説明していくことにしよう。一口に統計といっても、われわれはいろいろな意味で用語を用いている。たとえば、

1．データの集まりそのもの

2．データを整理、分類する手法

3．平均や標準偏差とか、データの集まりから計算された数字

などと大きく分けることができよう。ここでいう統計学では主として、

- どのようにデータを集めたらよいか
- どういう解析法をおこなうか
- データからどの程度のことがいえるのか

を学ぶことが目的である。そしてこれらの解説に必要な数学的理論として確率論を導入する。

古代の中国やエジプトの王朝でも人口を取り扱うための統計データの集計が行われていた。Yao 王朝では紀元前２２３８年の人口調査の記録があり，ペルシャ，エジプトにも残されている。現代は統計法という法律が制定されている。統計の真実性の確保，統計調査の重複を除去，統計体系の整備，統計制度の改善を目的とする法律（1947 年）である。重要な統計を指定統計として公示される。また統計には欠くことのできない集計計算にともない，コンピュータの高速化，ネットワークの高度化も現代の統計の発達の大切な重要な要因として，認識しておかねばならない。

現代社会の天気予報では、地域の気象データを通信回線で刻々と集計し、人口衛星からの雲の状況をもとにして、できるだけ正確な予報をと進展されてきた。しかし、その情報がありながら、はたして傘をもって行こうか行くまいか迷う場合がある。傘をもたずに行って濡れて帰るのも嫌だし、雨が降らないのに傘を持ち歩く煩わしさも好きではないだろう。ここでわれわれは知らずに、ひとつのゲームに参加をしている。大げさに言えば、いわゆる、賭けをしているのである。空が晴れていて雲ひとつなく、雨の降る気配もなく、いままでの経験的状況から降雨の可能性が少ないのに、わざわざ傘を持って行く必要はまったくない。雨が降る降らないという結果には、確率変動が伴い、われわれは、その可能性をまわりの情報から判断しているのである。

ある事柄を決定するには、必ずそれによって得るものと失うものを比較し、結果のおこる可能性にもとずいて正しい判断を下す。これを考慮しない判断は合理的行動とは言えない。日常生活では絶えず出会う簡単

な事柄について、比較はいつしか習慣的に無意識におこなわれている。傘の問題ならば、迎えを頼んだり、タクシーを使えば済むことである。もし判断によって得るものと失うものがとてつもなく大きい場合にはどうするだろうか。比較している 2 つの結果の可能性について、危険を避けるには、より正確な知識を得ることである。不偏性をもった正確さには、客観性をもたなければならない。そのためには、数字を用い、対象を数字で表現することが必要である。

天気予報の例では降雨確率で表現している。しかし、いかにして対象とする状況を数字にできるか、いわゆる、計るということをしなければならない。自治体にとっては人口動態に応じて、小学校、中学校を建てるか建てないかという判断を下す場合には、正確な人口構成数をつかんでいなければならない。ここでこのような数量的事実を得るために、計るという操作、つまり、統計調査が行われ、このデータによる予測がおこなわれる。このように”不確実なもとでの決定”は、正しい情報にもとづく状態の認識が必要である。そして、どういう前提条件によって、しかるべき判断とその結論を得ているのか理解しなければならない。

2　いろいろな概念

一般教養として「統計学」を学習するうえで、またさまざまな専門分野において統計学の知識を応用していくために、必要な基礎的テーマをあげておく。いわゆる数理統計学において学ぶべきことの概要を挙げるとつぎのようになろう。

確率と確率分布

統計学を理解するために必要な言語、あるいは理論展開していく道具として、確率の概念は必須のものである。確率とは不確実な現象をそのまま記述するよう数値で与えられた尺度を意味している。この尺度に基づいて、集団を数学的に述べることができるよう、確率変数とか確率分布が定められる。偶然変動による結果となる標本を表すには、確率変数が用いられ、もとの集団は分布という形で述べられる。とくに用いる分布は、２項分布と正規分布が基本的であり、これらの分布から派生される標本分布として、t-分布、カイ２乗分布、Ｆ分布を用いて推定や検定が行なわれる。

統計データと標本調査

データはそれぞれ固有の性質から、量的属性と質的属性に分類できる。さらに量的なものは、離散型と連続型におおきく分けられる。前者は 2 項分布、後者は正規分布が代表的な分布であるが、日常に表れる分布はこのような名前のない一般形をしているものは非常に多い。質的属性は数値として得られるものではないから、われわれが計算し易いよう適当に数量化する処理を施すことが多い。データを得るにはまず、統計調査をしなければならない。どのようにしてデータを集めるかという調査の設計を考え、全数調査か標本調査かを決める。標本調査となれば、標本抽出をする。対象とする集団を統計用語で母集団、標本の個数を標本の大きさとよぶ。もし大きな母集団から、効率よく標本抽出をするための一つの工夫として、地域や階級などの補助情報を用いる。

記述統計

統計調査の結果得られたデータは、単に数値などの羅列にしか過ぎないが、この観測値の集りを要約した情報として整理をすることが記述統計とよばれるものである。具体的には、この集まりを 1 つの数値でまとめたり、いろいろな見易いグラフで表す工夫をしている。平均とか分散は典型的な位置やひろがり具合を表すための代表値である。また最近では探索的データ解析というデータのいろいろな記述を通して全体の構造を見出すことが行なわれている。

推測統計

統計調査のデータを記述統計によって集約してみると、そこには新しい発見が見出せるかもしれない。しかし実際、未知の状態を理論として裏付し、認識するには、真の状態を推定したり、いくつかの考えられるものから真の状態を判断する検定がおこなわれる。このように観測値から推定や予測を行い、新しい状況に対して、推測する基礎としてもちいる理論的方法を推測統計とよぶ。母集団から抽出された標本にもとづいた母数（パラメータ）の推定、信頼区間や仮説の検定が具体的な方法である。最近では統計的決定理論として推測問題は発展している。大まかに統計学は記述統計と推測統計とに、上で述べた手法の違いによって分類される。

予測のための解析

統計理論として1次元のデータ、すなわち1変量の理論が簡単であるが、実際の分野に統計を応用しようとすると、多変量の理論が多いし、現状は複雑であってどうしても必要となろう。一般に多変量解析とよぶ多くの変量間の相互関連を分析する手法がある。いく組かの標本を抽出して、ある特性に関して有意に異なっているか否かを判断するために、標本の分散を分けて分析する分散分析や、変量の間にモデルとする回帰式をたてて、ある目的とする変数の測定値の変動が、モデルによって説明する変数によって充分満足のいくものであるかどうか、現象の理解や目的変数の予測をすることが回帰分析であり、現代では、心理学、社会学、政治学、生物学、農学、医学、工学など多くの分野で用いられている。