

データのまとめ方

1 統計の調査

統計調査をおこなって記述されたデータの結果は、大まかに、質的属性と量的属性の2種に分けることができる。質的属性とは、世帯主の職業とか有権者の支持政党など、調査結果が数値ではなく、いくつかの項目にいずれか属するとか属さないという形で与えられる。質的属性の分類項目を **カテゴリー (category)** とよぶ。また属性をカテゴリーに分類するとき、予めなんらの方法により、数量化しておいたほうが統計解析を施し易いことが多い。しかし複雑な多面的現象や微妙な心理的特性、知能や好み具合を数量化することは非常に難しいことである。これに対して、統計調査の観測結果が数値により記録される場合、そのデータを量的データという。量的データには、 $0, 1, 2, 3, \dots$ など自然数や整数を結果としてとり得る計数データ (count data) と、小数 (あるいは分数) の値をとり得ると考えられる計量データ (measure data) に分類される。長さ、重さ、温度など数直線上の目盛り、実数を測定するものも計量データである。後の確率的取り扱いと対応させるためには、前者を **離散型データ (discrete data)**、そして後者を **連続型データ (continuous data)** と呼んだほうが都合がよい。いずれの2つも、これらは計量データであり、理論上、測定を精密にすることにより、実数軸上の任意の値をとり得るが、測定値として記録する場合、離散化が施される。野球選手の打率は、小数第4位で丸めて、 $0.263, 0.320$ などとなり、また身長は mm の単位で、体重は kg で測定し、 $100g$ 単位で丸めるのが普通に行われている。測定値が離散的になっていても、本来の変数が連続的に変化すると考えられるならば、それは連続的な変数である。連続的な変数の測定値は離散化されて記録されるが、元々の離散型データとは異なる状況にある。統計解析をおこなうために、調査、観測をした結果、得られた個体の特性を表す数値を変数という。

連続型変数 数値が小数で表される、連続的な値をとる

離散型変数 整数や自然数など、とびとびの値をとる

統計データ	
質的データ	カテゴリー (項目) で分類
量的データ	計数データ 離散型変数の観測値
	計量データ 連続型変数の観測値

2 標本の抽出

標本の抽出には統計調査の規模、正確度、また必要性、興味の対象からいろいろな方法がとられる。いわゆる公平性、偏りのなさを求め、全体をなるべく正確に把握するには、調査の結果が客観的に受け入れられるよう、無作為抽出が行われる。対象とする母集団にきわめて近い特徴をもつような標本を抽出するため行われるものが確率標本抽出法である。これには関して

- 単純無作為抽出【たんじゅんむさくいちゅうしゅつほう】
- 2 段抽出【2 だんちゅうしゅつほう】
- 層化 (層別) 抽出【そうか (そうべつ) ちゅうしゅつほう】

を説明しよう。統計調査の結論は社会的に大きな影響を与える。政治的、経済的な政策決定はもちろんのこと、市場調査、将来予測、需要の調査などは企業にとって社運を決定することもありえる重要なことである。国勢調査には統計法として定められた法律の義務があり、将来の社会政策や人口予測、経済活動などには極めて重要なことである。ましてや意図しないあるいは気づかれない誤謬を犯さぬよう、騙されぬようきちんとした統計知識の判断知識をもたなければならないと思う。

3 単純無作為抽出法

母集団のすべての構成要素に番号をつけておき、乱数表や計算機の発生する乱数（擬似乱数）などにより、構成要素を同じ確からしさで抽出する。母集団の大きさが大であると抽出は困難となるが、理論的な解析は容易にできる。 N 個からなる有限母集団から n 個の標本を無作為に抽出するのは次のように行う。

- 1 母集団の構成要素の全てに通し番号をつける。
- 2 N 個の番号から n 個の番号をでたらめに選ぶ。
- 3 選ばれた番号の標本を実際に調査する（抽出する）。

ここで、 N 個からなる有限母集団から n 個の標本を無作為に抽出するしかたは、作意や主観など人為的要素の混入を防ぐために道具をもちいる。乱数サイという、正 20 面体のサイコロを投げて番号を決めたり、乱数表を用いたりする。多くの場合はコンピュータで擬似乱数を発生させる。乱数表から番号を選ぶとき、たとえば 735 人の集団から、5 人を選ぶには、3桁の乱数（000 から 999）をつくって、このうちから番号 1 から番号 735 に対するもの 5 個を選ぶが、もし該当しない番号には除外していけばよい。あるいは計算プログラムにより一定の範囲内（番号 1 から 735 まで）の擬似一様乱数を 5 個作り出せばよい。また宝くじなどの抽選では、回転する円盤に矢を射ることで抽出する状況をテレビで見たことがあるであろう。

4 2 段抽出法

対象とする母集団全体をある属性によって、いくつかの部分集団に分ける。これを層とよび、各々の層から無作為に標本を抽出する方法である。各層からはどのくらい抽出したらよいか、標本の大きさを決定する問題には、比例割当法やネイマンの最適割り当て法が知られている。また母集団をあらかじめいくつかのブロック（グループ）に分けて抽出して、選んだブロックすべての標本とする方法を集落（ブロック）抽出法という。集落抽出法によって抽出した標本集落の中に、構成単位が多くあるとき、その全部を標本としないで、再抽出を行って標本とする方法が 2 段抽出法とよばれる。この場合、初めに集落を抽出する事を 1 段目の抽出又は第 1 次抽出、集落の中での抽出を 2 段目の抽出又は第 2 次抽出という。このような方法は母集団の構成要素から個体を直接抽出するより、効率性や経済性、実行可能性あるいは調査のもつ性格などを考慮してそれぞれ工夫がなされている。

一般に全国規模の大がかりな調査では、前述のような単純無作為抽出法は作業量が膨大なものになる。そこでまず、市町村を抽出単位として無作為抽出する。つぎに選ばれた市町村の中でそれぞれ前述の方法で個体を抽出する。これを多段抽出法という。抽出は何段階でもよい。例えば、最初に都道府県を抽出し、市町村を抽出し、地区を抽出し、その後で個体を抽出するなど。ただし単純無作為抽出法などに比べると精度が落ちてしまう。

ある都市から 10 人を選ぶには、住民台帳などをもとにこの都市の住所により、区、町別にあらかじめブロックに分けて、これから第 1 段階の抽出をおこない、つぎに選ばれた区や町から番号付けておいた人を選ぶ。多くは人口に比例した標本の大きさで選ぶ確率比例法が用いられる。

例えば、家計調査を行うのに、初めに調査区を抽出し、次に、抽出された調査区の中から世帯を抽出するという方法が 2 段抽出法の一例である。2 段抽出法においては、各段で抽出単位が異なるのが特徴である。すなわち、第 1 次抽出単位は第 2 次抽出単位の集落である。しかし、取り扱う項目は各段共通である。この点、2 相抽出法と対照的である。平均は集団に属するすべての単位を合計し、単位数で割ったもの、いわゆる重みつき平均で集団の平均が計算される。

5 層化抽出法

ある母集団を構成するブロック（母集団に対応しているという意味）を分割して、部分母集団となる幾つかのブロックを作り、各ブロックから標本を抽出するとき、各ブロックに対応する部分母集団を層という。層を作ることを層化、各層から標本を抽出することによって全標本を得るとき、その抽出方法を層化抽出法という。層化は抽出単位に対して行われる。抽出は、原則として、層ごとに独立で行われる。層ごとの抽出は、単位無作為抽出法でも系統抽出法（系統抽出法とはつぎの抽出法：母集団の全構成要素に通し番号をつける。はじめの一つの標本だけは乱数表などでランダムに選ぶ。それ以降の標本はこの数字から始めて一定間隔で抽出する。この方法は、標本抽出枠が非常に大きく、単純無作為抽出法が難しい場合に用いられる。ただし番号の付け方に一定の周期性があるときには偏りが生じるので注意が必要である）でもよい。また、ある層で単純無作為抽出法を用い、別の層で系統抽出法を用いてもよい。層化の効果は調査項目ごとに評価される。ある項目に対して層化を効果的にするためには、単なる枠の分割でなく、その項目について同質なものを同じ枠にする、という方針で層を作らなければならない。これは、ちょうど、集落抽出法において効果的な集落を作るために同質なものを別の集落に、異質なものを同じ集落にするという方針と逆である。たとえば、ある都市の事業所・企業統計に対して調査をすることを考えよう。抽出に先立って、企業を種別に分け、農業・鉱業・製造業・サービス業・運輸・通信業などのような等質のグループに分け、それぞれのグループから標本を抽出する。このときの、等質な各グループを層、層に分けることを層別化という。景気の状態を把握するには、標本を無作為に抽出した場合、母集団の構成がそのまま標本に反映するとは限らないから、このような工夫をおこなう。

6 標本の割り当ての大きさ

層別抽出法には 最適割当法 と 比例割当法 がある。よく用いられるのは比例割当法であり、必要な標本の大きさを n 、各層の大きさを N_i ($i = 1, 2, \dots, k$)、母集団の大きさを N としたとき、各層から抽出する標本の大きさ n_i は、 $n_i = n \frac{N_i}{N}$ とする。

統計データの参考資料

統計資料はインターネットでたくさん入手できます。政府統計の総合窓口としてポータルサイト¹として <http://portal.stat.go.jp/> が代表的なものです。また総務省統計局は <http://www.stat.go.jp/index.htm> は日本の統計を行政の立場から国勢調査などを実施し、管理する重要な役割をもっているところです。統計データがたくさんあるので一度は WEB 訪問してみたらいいとおもいます。

¹ インターネットの入り口となる巨大な Web サイト。検索エンジンやリンク集を核として、ニュースや株価などの情報提供サービス、ブラウザから利用できる Web メールサービス、電子掲示板、チャットなど、ユーザがインターネットで必要とする機能をすべて無料で提供して利用者数を増やし、広告や電子商取引仲介サービスなどで収入を得るサイトのことをいう。出典：<http://e-words.jp/IT> 用語辞典