…… いんふぉるむ（第3回）……

＜サンプリングについて＞

調査の目的が決まれば、その目的にあったサンプリングが行なわれる。しかし、目的にあったサンプリングといっても、「２０代の政治意識についての調査」や「婦人の意識調査」のように容易に対象が設定される場合ならよいがそうばかりとは限らない。「短波放送の聴取者」や「年金制度を検討中の事業所」などを対象とする調査の場合には、調査結果の有効性と調査の効率を念頭におきながら対象の設定を考えていかなければならない。さらに、対象の設定についても唯一の「正解」があるわけではない。調査の目的を達成することができれば、合理的な理由がある限り幾通りかの標本設計が考えられるのである。そして、そのサンプリングによって調査結果は様々なものとなる。このように、サンプリングは調査の性質そのものを決定する重要なものであり調査の要である。このサンプリングに対する考えが理解されていないと、いかに優れた調査票を作成しても、いかに優秀な調査員によって調査が実施されても調査自体が無意味なものになってしまう。

今回はこのサンプリングについてとりあげる。前半はサンプリングに関する基本的な事柄および理論を紹介し、後半（次号）で実際の調査におけるいくつかの問題点を検討する。

標本抽出が必要な場合

よく、街を歩いている人に聞くアンケートというものがある。行きあたりばったりに人をつかまえて一定の数を確保するやり方である。調査というものは、どのようなものでも、常にある条件の下で行なわれるものであり、日時、場所、状況等が詳しく説明されているならばこのようなアンケートもすべて無意味であるとはいえない。しかし、問題なのはその調査によって何を知りたいか、である。もし、このアンケートによってある新製品の売れゆきを探ろうとするならば調査の対象者をこのようにして選ぶことには問題がある。たまたまそこを通りかかった人というのは、場所や時間によってかなりの偏りがあると考えられる。それなのに、調査によって知りたいことが市場におけるある商品の売れゆきであるとすれば、市場というワクと調査の対象ワクとの間にズレが生じる。これでは目的にあった正しい調査とはいえない。

このような場合には、市場を構成している人の意見が正確に反映されるように調査の対象を設定する必要がある。それでは、市場を構成している人の意見が正確に反映するとはどういうことであろうか。その商品によるが、例えば化粧品などのように女性を対象とするものであれば、購入者として考えられる××才から××才までの女性の意見が正確にとらえられる、ということである。仮に、東京を中心とした一都三県に住むこれらの女性のうち４０万人がこの化粧品を買いたいと思っていることがわかれば、それを目安として営業方針も立てられるし製造計画も立つ。そして、その数字が事実に近いほどロスが少なくなる。つまり、何百万かの女性全員の購入意向がつかめれば良いのである。しかし、実際にはそんなに大勢の意見をいちいち調べていたら時間もかかるし労力（費用）もかかる。これでは採算があわない。

確かに全体を正確に把握することは大事であり、そのためには対象となる全部を知ることが望ましい。しかし、時間的にも経済的にも制約を受けていればその範囲内でできる限り正確な情報を得るべきであろう。そこで標本抽出による調査が必要となってくる。全体の意見を探るのに一部の標本をとって調査をすること、これが標本調査である。

ところで、標本調査というのは確かに便利で有効な考え方ではあるが、いかなる調査も標本調査とは限らない。例えば、ある高校生の３年生の進路希望を調べたいというときに、５００人の生徒の中から１００人だけ選んで調査をする場合を考えてみよう。調査の主題が生徒の進路希望というのであれば、その調査結果は生徒ひとりひとりの進路相談の資料となるはずである。もしそうであるならば、５００人の中から１００人だけ選んで全体の傾向を探ったところで何の意味もない。調査の目的を考えたうえで標本調査にすべきかどうかを決めるべきである。

さて、標本調査を行なうとして、次には全体の中からどのようにして一部の標本を選び出すかが問題となる。選ばれた標本が行きあたりばったりのアンケートと同じであってはなんの意味もない。この標本の選び方についての考え方、標本抽出の方法、さらに実際の抽出の手順といったところを以下で紹介してゆく。

科学的なサンプリング調査の登場

１９３６年、アメリカの大統領選挙においてリタラリー・ダイジェスト社は全国の10,000,000人に対して郵送調査を実施した。10,000,000人という人数は、当時のアメリカの３世帯に１件の割合であった。対象者は電話帳あるいは自家用車所有者リストからピックアップされた。回収された2,376,523票（回収率２３.８％）の調査結果によると、共和党の候補であるランドン（５７％）が民主党候補ルーズベルト（４３％）をおさえて勝利すると予想された。ところが、実際はルーズベルトが６３％、ランドンが３８％でルーズベルトの大勝に終わった。この時、リタラリー・ダイジェスト社とは別にそれまで実験的にいくつかの世論調査を行なっていたジョージ・ギャラップは確率論（the law of probability）による標本抽出をもとにはるかに少ない標本数でこの結果を予測した。ここで明らかにされたことは、世論調査において重要なことは調査の対象となる母集団全体の姿が正確に反映されることであり、そのための科学的なサンプリングが必要であるということであった。

彼のサンプリング理論の基本は「偏らず、平等に」ということである。彼によれば、リタラリー・ダイジェスト社のサンプリングは電話や自家用車を所有している人であるから生活水準がある程度高く、この生活水準の程度が実際の投票行動に少なからずの影響を与えるというのである。そこで彼は、サンプリングは数ではなく選び方が問題なのだ、ということを主張した。

「数は問題ではない」という主張を説明するときに、彼はナベの中のスープの話を持ち出す。つまり、ナベに入ったスープの味をみたいときにはスープを全部飲む必要はなく、スプーン一杯だけをとり出して味見をすればよい。ただし、この場合はナベのスープがよくかきまぜられたものでなくてはならない。よくかきまぜられたものであれば、スプーン一杯でもスープ皿一杯でも味は変わらない。彼のサンプリングの理論はこのようなものであった。いい換えれば、調査の対象となる母集団と同じ性質を持った対象者集団の設定ができれば「数は問題ではない」ということになる。

しかし、勿論、１億人の意見を知るのに５人や１０人の意見を聞けば済むというものではない。ある一定の数で全体の傾向を知ろうとする場合には、必らずその場合の精度というものが問題になる。一定の数（一定の精度）が確保されればそれ以上数を増やすことはさほど重要ではない、というのが彼の主張するところである。また、「偏らず、平等に」とは母集団を構成しているひとりひとりが等しく対象者に選ばれる機会を持つことである。例えていうならば、母集団の数だけのカードを裏返しにしてよくかきまぜたうえで何枚かをぬき出すという考えである。こうすれば誰がぬき出されるかまったくわからないし、特定の人を恣意的にぬき出したり排除したりすることができなくなる。もっとも、これは単なる考え方にすぎず実際の抽出の手順はかなり合理的に行なわれる。抽出の手順については後述する。

標本抽出の方法

標本抽出にはいくつかの方法がある。代表的なものを示すと①単純無作為抽出法　②等間隔抽出法　③多段抽出法　④層別抽出法　といったところがある。

①単純無作為抽出法：これは前述したとおり、カードを裏返しにしてよくかきまぜたうえで何枚かをぬき出すという方法である。どのカードも抽出される確率が等しくなる。

②等間隔抽出法：ある数字を決め、集団の中からその数字番目ごとに順次ぬき出して行く方法である。例えば集団の数が1,０００でこの中から100の標本をとりたい場合は集団の各構成単位に１から1,０００までの連番をつけ最初の対象を任意にとり、1,０００からその対象の番号を引いた数を１００で割る。例えば最初の対象が８９だとすると、（1,000－８９）÷１００＝9.１１となり、その後９番目おきに順次とっていけば集団全体にわたって効率よく抽出することができる。勿論、実際には連番をつける必要はなく、集団の数さえわかれば計算して数えていくだけでよい。ただし、この方法だと抽出台帳に一定のオーダーがあった場合、特定の対象だけが抽出され偏った集まりが作られてしまう危険性がある。例えば、夫婦所帯の団地などの場合は世帯主だけとか配偶者だけとかになってしまう。

③多段抽出法：母集団が大きな範囲で単純無作為抽出や等間隔抽出をすると各対象のバラツキが大きくなり、実際に調査員が訪問するような調査方法では実査が極めて困難となる。そこで、第１次抽出単位として実際に調査が可能な範囲をぬき出し、次にそこから対象を抽出するという方法である。この場合は２段抽出法といえるが、この抽出次元を何段にも設定することができるから一般的には多段抽出法といわれている。

④層別抽出法：母集団が予めある特性をもったグループ（層）に分けられているとき、これらのグループ別にサンプルを抽出する方法である。このねらいは、共通の特性をもったグループ内では分散（バラツキ度）を小さくし、逆にグループ間の分散を大きくしてサンプリングの精度を高めることにある。ここでは、どのように層別するかがその調査の質を左右することになる。

これ以外にも、最初に大サンプルを抽出し、それらを層別して最終標本を確定してゆく標本層別抽出法や、③２段抽出法と④層別抽出法を合体させた層別２段抽出法などの手法がある。

サンプリングの実際

日本国民2,０００人を対象とした世論調査の場合をとりあげ実際の抽出手順を紹介しよう。ここでは前段の最後に触れた層別２段抽出法という方法が用いられる。

まず最初に、2,０００人を地域別・都市規模別にそれぞれの人口数によって比例配分をする。つまり、地域別・都市規模別に層化する。ここで、関東の大都市部では２１０サンプル、九州の郡部では８５サンプルという具合にそれぞれの地域別・都市規模別にサンプル数が決められる。そして、九州の郡部（九州全域にわたる、いわゆる「市」以外の全町村）から８５サンプルをバラバラに選ぶとなると実査が大変であるから、まず第一次抽出単位として調査地点を抽出しようというわけである。その時、一調査地点内に１５人前後の対象者が当たるように調査地点数を決めるが、この際の１５人前後というのは定められた調査期間内にひとりの調査員が調査を完了し得ると考えられている経験的な数字である。このようにして東京の大都市部では１５地点、九州の郡部では６地点という地点数が抽出される。具体的に東京都千代田区永田町以下１５地点、鹿児島県薩摩郡薩摩町以下６地点が決まると、その調査地点の中から定められただけの対象者個人を抽出してゆく。個人を抽出する際は、その調査地点を管轄する役所にある住民台帳を利用することが多い。

ここでは２つの数字が与えられる。ひとつの数字はその調査地点の中で最初に選び出される人を決める数字である。まず住民台帳の中から調査地点の部分を取り出し、対象となる人（例えば２０歳以上の男女とか１８歳以上の女性など）を数えていってその数字番目に当たった人を第一対象者とする。第２対象者以降は、選ばれた対象者から数えてもうひとつの数字番目に当たった人を次々に抽出してゆく。この最初の数字は乱数表からとり出した数字であり、２番目の数字は調査地点から隈無く対象者が選び出されるように計算された数字である。

以上の抽出においては、調査地点として考えられるすべての地区単位が実際に調査地点として抽出される機会が均等であること、さらにその中のひとつの調査地点内に居住している個人ひとりひとりが対象者として抽出される機会が均等であることが原則とされる。換言すれば、日本国民の誰もが調査対象者として選ばれる確率が等しいことがこの標本抽出の基本なのである。

標本誤差

全体の中から一部だけの人の意見を聞くのであるから、真の全体の意見とはモノが違うわけである。このモノの違い、換言すれば本物との距離を示すものに標本誤差というものがある。

ある質問に対して「ハイ」と答えた人が８０％、「イイエ」と答えた人が２０％いるとする。このときの標本誤差が５％だとすると、「ハイ」と答えた人の８０％という比率は７５％から８５％までの広がりを持っていることになる。標本調査による調査結果は常にこの標本誤差を念頭に置いて読みとる必要がある。例えば、ある世論調査の結果、ある質問に対する答えが「ハイ」５３％、「イイエ」４７％であるとする。この時、「世論調査の結果はハイがイイエよりも６％多い」ということになる。しかし、標本誤差が４％だとすると、「ハイ」は４９％から５７％、「イイエ」は４３％から５１％という広がりを持っているから真の全体の意見はどちらが過半数なのかはわからない。調査結果をもとに母集団のことを考えるとき、標本誤差はこのように真の全体の姿を把握するうえでの重要な指標となる。しかし、それでもなお、「世論調査の結果はハイがイイエよりも６％多い」という事実に変わりはない。調査結果は標本誤差を内包したうえでのひとつの事実であり、これをもとに解釈される真の全体の姿とは別個の

ものだからである。

調査結果と母集団が標本誤差を仲介者としてどのような関係にあるかは上述したとおりであるが、標本誤差そのものはどのような意味を持った数字なのか。

まず、標本誤差に関係してくるものとして母集団数、標本数、回答の出方があげられる。常識的に考えてもわかるように、母集団数に対して標本数の割合が多いほど誤差は少なくなる。また、回答の出方がひとつの方向に集中していれば、全体としての傾向が探りやすいということであるから、この場合の誤差が回答が分散した場合よりも小さいことは肯けるであろう。標本誤差とは、要はこのような意味を持った数字である。参考のために、これを数式に表わすと次のようになる。

$$b = 2\sqrt{\frac{N-n}{N-1} \cdot \frac{P(1-p)}{n}}$$

b＝標本誤差
N＝母集団数
n＝標本数
P＝回答率

この数式を用いると、母集団の人数が5万人のときに標本誤差を5％以内におさえようとすると最低いくつの標本数が必要になるかが計算できる。回答率を仮に「ハイ」60％、「イイエ」40％とする。数式に与えられた数字を代入してnを求めると、n＝381となる。同様に標本誤差を2％以内におさえようとするときのnの値を求めるとn＝2,290となる。

標本誤差2％というのは「ハイ」と「イイエ」がそれぞれ47％―53％のときにものがいえるということであり、逆に標本誤差5％では45％―55％でもものがいえないということになる。調査の設計の際にはこのような数字の意味を考慮しながらサンプル数を検討することもある。

以上、サンプリングの基本的な事柄と理論について紹介した。次号では現在のサンプリングをめぐるいくつかの問題について検討する。

（編集部）